REMOTE SENSING AND MODELING OF *VIBRIO* BACTERIA IN THE CHESAPEAKE BAY

by
Erin A. Urquhart

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland
April, 2014

ABSTRACT

Estuaries and coastal waters are dynamic environments, subject to mixing processes that produce high temporal and spatial variability in water properties relevant to water quality and ecology. These environments are also increasingly vulnerable to adverse environmental and biological change under pressures of human population growth, sea level rise, and climate change. In coastal regions such as the Chesapeake Bay, for example, it has been suggested that the occurrence of *Vibrio* spp. bacteria is increasing throughout the near shore environments. As environmental conditions continue to change in poorly characterized and unpredicted ways, there is a need for more advanced and spatially complete coastal monitoring networks. The objective of this dissertation focused on using environmental predictors to develop a *Vibrio spp.* bacteria estimation method for the Chesapeake Bay with near-real time forecasting potential. This dissertation work has involved development of a satellite-derived surface salinity product generalizable to the Chesapeake Bay, geospatial interpolation of remotely sensed surface water temperature and salinity, comparison of satellite-derived and hydrodynamically modeled estimates of environmental predictors relevant to *Vibrio* occurrence, development and validation of *Vibrio* spp. likelihood and abundance models, and lastly sensitivity in modeled response of *Vibrio* to observed and projected temperature and salinity changes in the Chesapeake Bay. The intended outcome of this research is to use the information of the satellite, interpolation, and modeled products to inform operational and public health risk models for *Vibrio* spp. in shellfish and recreational waters in the Chesapeake Bay. Though *Vibrio* spp. does not pose a serious health threat in the Chesapeake Bay, using the Chesapeake Bay as a model "test bed" has provided valuable model information, and

help quantifying model uncertainty that will later be extended to other regions of the world significantly aiding in the prediction and, potentially, prevention of *Vibrio* outbreaks.

Advisor:

Ben F. Zaitchik, PhD, *Department of Earth and Planetary Sciences*

Thesis Readers:

Darryn W. Waugh, PhD, *Department of Earth and Planetary Sciences*

ACKNOWLEDGEMENTS

I would like to thank my advisory committee: Ben Zaitchik, Darryn Waugh, and Carlos del Castillo. First and foremost, Ben has been an incredible academic mentor, editor, teacher and friend over the past five years, allowing me to pursue my own interdisciplinary research interests, however far "the little swimmy things" were from his own area of research. I admire Ben's spirit and compassion for his scientific research, his eagerness for interdisciplinary collaborations, and his unbelievable patience for students such as myself. I would also like to acknowledge Darryn Waugh, who not only showed great faith in me by accepting me as an un-conventional "guinea pig" into the EPS department, but who also stood by my interdisciplinary research interests and goals throughout the entire duration of my study. Lastly, I would like to thank Carlos del Castillo for his undeniable enthusiasm in my research and help in making new academic connections in the field of coastal remote sensing. I am honored to join the ranks of Carlos's successful "grasshoppers". I cannot thank the three of them enough for their tolerance, humor, and support ranging from fieldwork logistics to everyday advice.

In addition I've been privileged to work with many invaluable collaborators on my research. Matthew Hoffman was instrumental in not only standing in as my personal math and MATLAB tutor, but also in providing much of the ChesROMS data that was employed in this dissertation. Seth Guikema offered me the statistical skills needed to quantitatively and qualitatively answer the underlying questions pertaining my research. Anand Gnanadesikan fostered excellent discussions and great contributions to the biological and physical aspects of the Chesapeake Bay research.

To Greg Henkes, Tiffany Smith, Sophie Lehmann, Holly Brown, Stephen Jeffress, Rebecca Kraft, and all of my other fellow EPS graduate students, I am thankful for the friendships that we have forged over many academic headaches and beer hours. I look forward to remaining lifelong colleagues.

This work would not have been possible if not for the financial support of the Department of Earth and Planetary Sciences, the JHU Global Water Program, and the NASA Headquarters Applied Sciences Division. I would also like to thank the Maryland Department of Natural Resources, Chesapeake Bay Program, and NASA GEO-CAPE field campaigns for allowing sample collection upon routine Chesapeake Bay cruises.

Lastly, I thank my friends and family for their consistent support and patience for my doctoral education. My parents Ken Urquhart, Linda and Phil Holland have offered continuous support and encouragement for my education, from elementary through graduate school. If it were not for their unrelenting questioning of my academic and career goals, I may have never been pushed to pursue my passion in the field of environmental health science. I'd like to thank my Baltimore friends particularly Sam Hanson, Stephanie Walters, Katherine Gorman, Andrew Klein, Ciaran Harman, John Berggren, Pamela Malo, Marissa Hildebrant, Brent Kim, Dan Jones, and Rose Smith for their fun and distractions from the world of academia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. CHAPTER 1: INTRODUCTION

1.1. *Vibrio* spp. bacteria

Vibrio spp. bacteria are a threat in many marine and estuarine ecosystems around the world (Baker-Austin et al., 2012; Deepanjali et al., 2005; Hendriksen et al., 2011; Cantet et al., 2013; Oberbeckmann et al., 2012). Halophilic *Vibrio* bacteria are natural inhabitants of aquatic environments and are dependent on optimal environmental and climatic conditions to persist. Numerous studies (Heidelberg et al., 2002; Jacobs et al., 2010; Wright et al., 1996; Louis et al., 2003; de Magny et al., 2009) in various regions of the world, along with extensive laboratory experiments, have shown that *Vibrio* bacteria follows a prominent seasonal cycle in estuarine environments attributed mainly to water temperature fluctuations, salinity changes, and primary productivity peaks, and coastal eutrophication (Figure 1.1). Upon exposure to the aquatic-environment adapted *Vibrio* bacteria (either through seafood consumption or direct water exposure), pathogenic strains infect the human host then subsequently shed off back into the environment. In the case of *Vibrio cholerae*, bacteria that have been shed off through waste are in a hyper-infectious pathogen state, which amplifies the disease outbreak to the subsequent host (Nelson et al., 2009).

Certain strains of *Vibrio* bacteria can cause human infections, which typically manifest as wound infections, gastroenteritis, or a syndrome known as primary septicemia (Morris and Black, 1985; Howard and Bennett, 1993). Transmission is predominately via ingestion of contaminated seafood or water or though direct bacteria penetration of wounds. *Vibrio cholerae* is the most well known member of the *Vibrio* genus, affecting

many developing countries that lack proper water sanitation resulting in approximately 3-5 million cases per year (WHO, 2013). *Vibrio vulnificus* illness is known to have one of the highest mortality rates of any foodborne illness and has recently become a WHO foodborne safety concern in many regions including: United States, New Zealand, Europe, Japan, and the Republic of Korea (WHO, 2005). High rates of *Vibrio parahaemolyticus* illness is frequently found in parts of Asia, Europe, South America, and the United States due to consumption of raw seafood. Between 1997 and 1998 there were approximately 700 human illnesses associated with *V. parahaemolyticus* in the United States due to the consumption of contaminated raw oysters (CDC, 1998,1999).

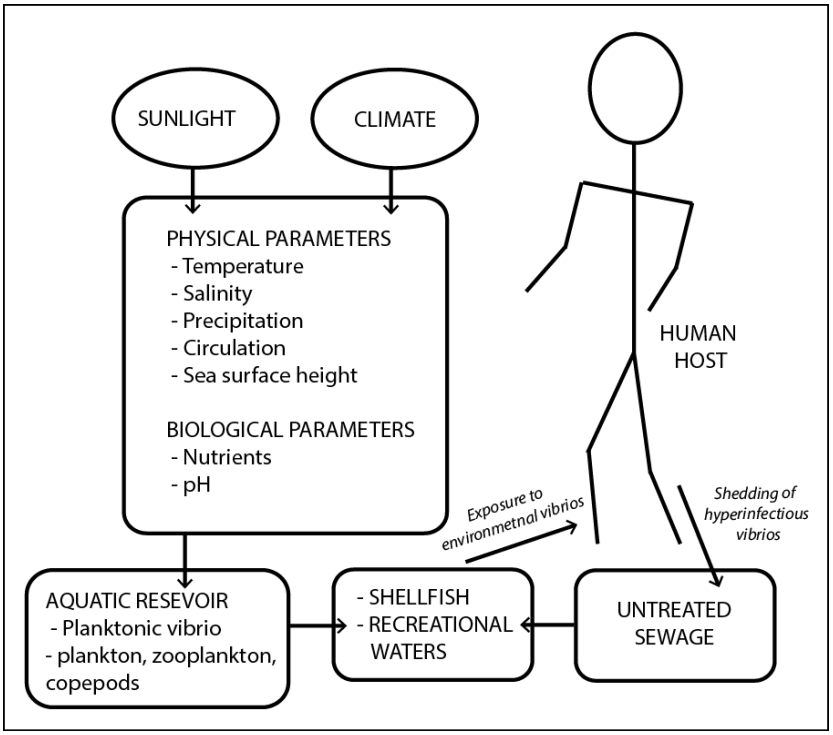

Figure 1.1 Lifecycle of *Vibrio* spp. in the environment (adapted from Nelson et al., 2009).

Whether it is negative environmental implications or public health manifestations around the globe, there is no scientific consensus for how future climatic variability and change will affect these marine microorganisms. Although there is considerable uncertainty in global sea surface warming, precipitation changes, and sea level rise due to climate change, is anticipated that these environmental changes could raise public health risk of *Vibrio* spp. bacteria throughout coastal regions worldwide. It is essential that before *Vibrio* outbreaks become a significant problem relative to geographical region, universally in terms of human health, the scientific community fully recognize the future of this coastal phenomenon. Hypothetical climate perturbations of the coastal system through use of forecast models, hindcast simulations, and remote sensing data can aid *Vibrio* spp. prediction research in the future.

1.2. Chesapeake Bay

The Chesapeake Bay estuary is the largest and most productive estuary in the United States with an area of 6,500 km^2, 300 km long, 48km wide, and a mean depth of 8.4m, (Marshall, 2006). The Chesapeake Bay extends from Havre de Grace, MD in the north to Norfolk, VA at the mouth of the Bay (Figure 1.2). Home to Baltimore, Washington DC, Annapolis and many other port cities, historically, the Chesapeake Bay has served as one of the country's main industrial shipping ports. An upper of 150 rivers and tributaries deliver streamflow into the Chesapeake Bay, with the Susquehanna River responsible for approximately 45 percent of the freshwater input to the Bay (Bankar et al, 2011). The Chesapeake Bay estuary has a strong north-to-south salinity gradient that includes

oligohaline (0-6[1]), mesohaline (6-18), and polyhaline (18-30) zones (Baird et al., 1989). Sea surface temperatures in the Bay range from local wintertime lows of -0.5°C to summertime highs of 31°C. The oligohaline upper Bay has a mean depth of 4.5m, the mesohaline middle Bay 10m, and the polyhaline lower Bay 9m, giving the overall Bay an average depth of 6.5m (22ft) (Baird et al., 1989; Figure 1.2). The physical transport regime of the Chesapeake Bay estuary follows the classical estuarine circulation model of partially mixed estuaries, in that it is characterized by a 2-layer gravitational circulation scheme. As salt water enters the mouth of the Bay along the eastern shore, there is a net up-estuary flow of water, which occurs below the pycnocline, and a complementary net down-estuary flow as the fresh surface water makes its way from the head to the mouth of the Chesapeake Bay (Pritchard, 1952).

Extensive estuarine wetlands surround the Chesapeake Bay and its numerous tributaries, offering shelter for a vast diversity of wildlife, migratory birds, and shellfish species (Water Encyclopedia, 2009). Unfortunately, due to natural and anthropogenic activities, there has been a drastic reduction in the overall health of the Chesapeake Bay marine environment (Chesapeake Bay Program, 2012; 2013). The monitoring of this "sub-par" marine environment, has received significant attention by state, academic, and national institutions. Needs for the improvement of monitoring techniques in the Bay can be attributed to various reasons including the increasing number of people who live by and depend on the coastal regions, as well as the huge economic dependence of local industry on the Bay ecosystem.

[1] In situ and estimated salinity values reported in this study use a standard unit less measure.

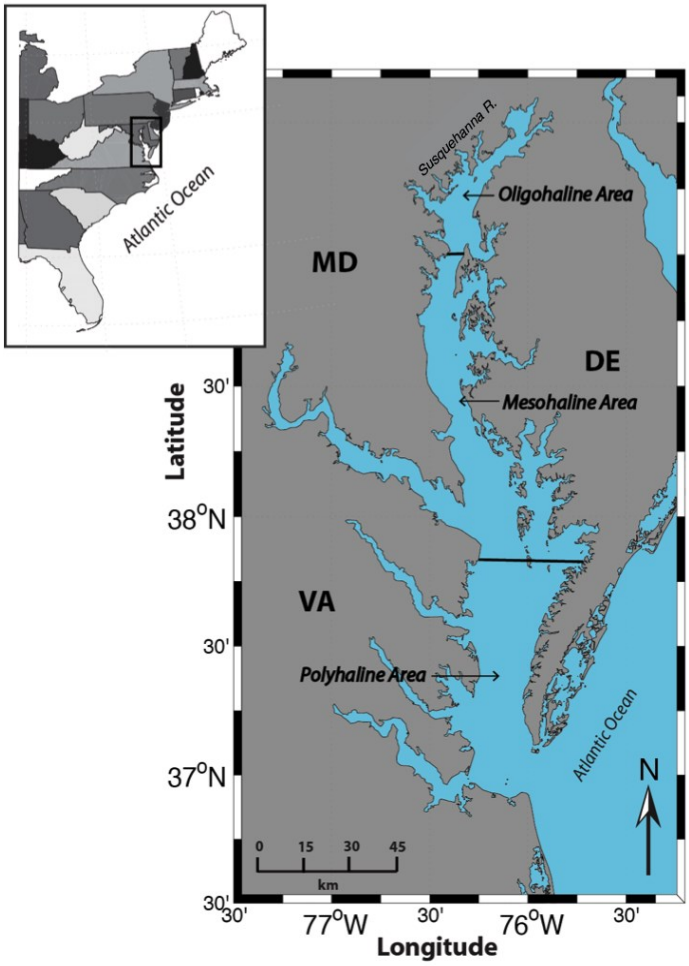

Figure 1.2 Map of Chesapeake Bay, with salinity regions identified.

In the dynamic estuarine environment of the Chesapeake Bay, near-surface salinity varies significantly in space and time. As absolute salinity and salinity gradients are central to many physical and ecological processes including the lifecycle of *Vibrio* spp. in the region, reliable and consistent salinity estimates are a priority for marine research and application communities. Satellite remote sensing has a great potential to meet this need, yet sensors and algorithms designed to monitor open ocean salinity are typically ill-suited for high resolution applications to coastlines and estuaries. In Chapter 2, I present results of multiple statistical models that predict daily surface salinity at 1 km resolution across

the Chesapeake Bay as a function of surface reflectance estimates from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS), onboard the Aqua platform. However, satellite data alone does not offer complete coverage of the estuary as satellite datasets are susceptible to incomplete coverage attributed mainly to cloud cover and coastline interference leading to data gaps that significantly hinder the broad application of satellite-informed predictions. In Chapter 3, the Chesapeake Bay estuary was used as a model "test bed" to which we applied the power of near real-time satellite-derived observations to the issue of water quality monitoring. To use remote sensing in support of spatially complete estimates of salinity and temperature in the Bay, I tested geospatial interpolation techniques as a method for filling gaps and minimizing errors in the satellite record.

1.3. *Vibrios* in the Chesapeake Bay

The natural marine microbiota of the Chesapeake Bay include several species of *Vibrio* spp. bacterium, some of which are harmful to the environment and opportunistic human pathogens (Colwell et al., 1977). While infrequent, the number of reported human *Vibrio* cases has nearly doubled in the past decade throughout the Chesapeake Bay region (Figure 1.3). First isolated from the Chesapeake Bay in 1970 by, *V. cholerae* was suggested to be autochthonous to the region (Colwell et al., 1977). In conjunction with the Center for Disease Control, the local health departments have compiled a ten-year record of cholera and non-cholera *Vibriosis* case reports for the Chesapeake Bay. Standard *Vibrio* case report data includes: bacterial species, source of infection (shellfish or recreational waters), time and location of exposure, pre-existing patient conditions, and status of infection. Mostly attributed to *Vibrio vulnificus*, the number of annual infections

in the region averages 60 cases per year (Maryland Department of Hygiene and Health, 2013; Virginia Department of Health, 2013). Furthermore, as of 2009, the Bay there has been 4 *Vibrio cholera* human cases (Figure 1.2). Though the ecological relationship between surface water temperature, salinity, and *Vibrio* bacteria in the Bay is known, there is no proven correlation between these environmental parameters and human *Vibrio* spp. infection. Previous *Vibrio* studies (Jacobs et al., 2010; Louis et al., 2003) in the region provide the likelihood of bacterial presence; but these studies do not characterize the risk of human outbreak.

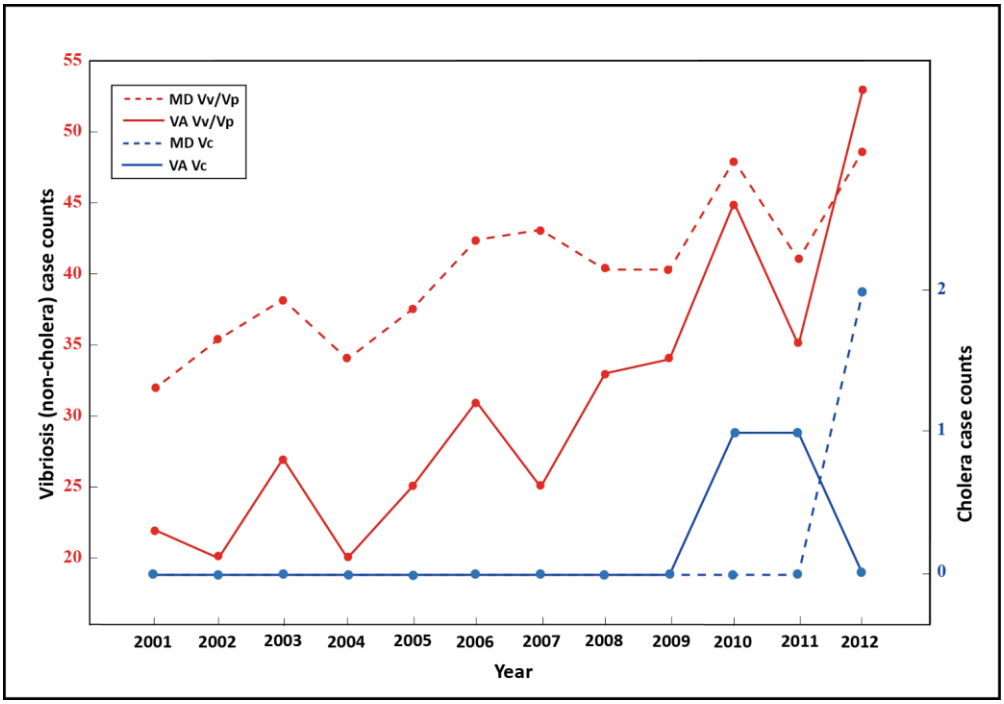

Figure 1.3 Reported *Vibrio* spp. cases in Maryland and Virginia (2001-2012; Maryland Department of Hygiene and Health, 2013; Virginia Department of Health, 2013).

Environmental *Vibrio* is autochthonous to estuaries, and depends on optimal environmental conditions to persist. Numerous studies (Colwell et al., 1977; Louis et al.,

2003) have shown that the spatial variation of *Vibrio* spp. abundance in the Chesapeake Bay is regulated by water temperature and relative salinity. The highest abundance of *V. cholerae*, both in the upper water column and in shellfish, occurs in cold waters (typically temperatures exceeding 15°C), and of low salinity values (2 to 15) (Constantin de Magny et al., 2009), while *V. vulnificus* is typically found in warmer waters (exceeding 20°C) with medium salinity (5 to 25) (Wright et al., 1996). That being said, a fairly strong spatial variation between *V. cholerae* and *V. Vulnificus* is found in the Bay. Ground based sampling (Louis et al., 2003; Jacobs et al., 2010) as well as empirically modeled predictions using in situ environmental data, have shown *V. cholerae* natural preference is towards the cold, low salinity, Upper Chesapeake and coastal tributaries, while *V. vulnificus* is more commonly found in the warmer, moderate salinity regions of the middle and lower Bay. The major drawback of all in situ empirical prediction models is that they fail to resolve the spatial and temporal scale often needed to develop accurate regional forecasting models. To address the in situ data limitations, the use of advanced techniques including satellite remote sensing and geospatial interpolation of surface water temperature and salinity, can be used to enhance the monitoring, prevention and mitigation of *Vibrio* spp. in the Chesapeake Bay. Furthermore, the environmental conditions associated with risk of *Vibrio* infection in coastal regions are poorly characterized, and as discussed in Chapter 5 the effect of climate change and variability on *Vibrio* spp. dynamics or risk of human *Vibrio* infection remains uncertain.

The published *Vibrio* spp. estimation models for Chesapeake Bay (Jacobs et al., 2010; Louis et al., 2003) are limited to the availability of in situ measurements of environmental predictor variables making it difficult to generalize predictions to the entire Bay.

Moreover, while these pre-existing models have been fairly successful in predicting *Vibrio* spp. prevalence in specific sampling areas of the Bay, they are constrained to likelihood by their ability to only predict prevalence. In Chapter 4 I present several empirical algorithms for predicting the probability of *Vibrio* spp. prevalence and abundance in the upper Chesapeake Bay. Since the risk of human infection is a function of *Vibrio* spp. concentration, extending available predictive models to provide concentration, in addition to presence/absence, advances the public health utility of the models significantly.

1.4. Dissertation Outline

This first chapter serves as an introduction to the issue of *Vibrio* spp. bacteria in the Chesapeake Bay, noting previous work and current data limitations. The second (Urquhart et al., 2012), third (Urquhart et al., 2013), fourth (Urquhart et al., 2014 (in review)), and fifth (Urquhart et al., 2014 (in review)) chapters represent published or in review scientific papers. Chapter 2 presents development of multiple statistical models capable of predicting surface water salinity across the Chesapeake Bay as a function of surface reflectance estimates from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS-Aqua). To use remote sensing in support of spatially complete estimates of salinity and temperature in the Chesapeake Bay, Chapter 3 outlines and evaluates two geospatial interpolation techniques as a method for filling gaps and minimizing errors in the satellite record. The fourth chapter describes development of empirical methods to model the likelihood of occurrence and abundance of *Vibrio* spp. in Chesapeake Bay. To explore the anticipated impact that future warming will have on *Vibrio* bacteria in the Chesapeake Bay, Chapter 5 compares three *V. vulnificus* likelihood

estimation methods and assesses each model's sensitivity to climatic variability and change within the region. Finally, the sixth chapter concludes the dissertation work.

2. CHAPTER 2: REMOTELY SENSED ESTIMATES OF SURFACE SALINITY IN THE CHESAPEAKE BAY: A STATISTICAL APPROACH[2]

ABSTRACT

In coastal and estuarine environments, near-surface salinity varies significantly in space and time. As absolute salinity and salinity gradients are central to many physical and ecological processes in these environments, reliable and consistent salinity estimates are a priority for marine research and application communities. Satellite remote sensing has a great potential to meet this need, yet sensors and algorithms designed to monitor open ocean salinity are typically ill-suited for high resolution applications to coastlines and estuaries. Here we present results of multiple statistical models that predict daily, gridded surface salinity at 1 km resolution across Chesapeake Bay as a function of level 2 surface reflectance estimates from the NASA Moderate Resolution Imaging Spectroradiometer (MODIS), onboard the Aqua platform. Eight statistical methods were tested and it was found that sea surface salinity can be accurately predicted via remote sensed products with an accuracy that is more than sufficient for many physical and ecological applications. For the best-performing statistical model, mean absolute error was 1.82 relative to mean Chesapeake Bay salinity of 16.5.

2.1. Introduction

Sea surface salinity plays a vital role in circulation patterns, influences the spatial distribution of many marine organisms, and affects seawater density in both coastal

[2] Urquhart E., Hoffman M., Zaitchik B, Guikema S., Geiger E. (2012) Remotely Sensed Estimates of Surface Salinity in the Chesapeake Bay: A Statistical Approach. *Remote Sensing of Environment*, 123: 522-531.

systems and open oceans. In coastal and estuarine environments, even small changes in salinity can greatly alter the transportation course and lifecycle of organisms and the status of the ecosystems they comprise (Baird et al., 1989). For this reason, salinity is a core input to ecological analyses and to operational models designed to monitor physical and biological processes in coastal environments. Advances in coastal remote sensing and computer modeling technology have led to several successful operational products that employ sea surface salinity. The National Atmospheric and Ocean CoastWatch Program provides a near real-time product for forecasting harmful algal blooms and predicting the likelihood of where sea nettles exist in the Chesapeake Bay. NOAA's forecasting models are accomplished by applying surface salinity and temperature estimated from a numerical hydrographic model (ChesROMS) to species-specific habitat models for the Bay (NOAA, 2010).

These applications point to the critical need for reliable, continuous, and spatially distributed estimates of salinity in coastal environments. In situ salinity measurements are a critical component of such monitoring efforts, but cost and logistics limit the temporal and spatial coverage of such measurements. The Chesapeake Bay Monitoring Program conducts routine bi-monthly water quality monitoring along the main-stem sections of Maryland and Virginia Bay waters. The monitoring program measures key components of the Bay ecosystem: habitat, living resources, pollutant inputs, and water quality. These monitoring efforts are used in both research and modeling of the Chesapeake Bay ecosystem (MDDNR, 2011). Both physical and biological processes in

coastal systems can occur on spatial and temporal scales that are not observed through monthly environmental sampling at designated sites and transects.

Satellite remote sensing offers the potential to estimate salinity across entire water bodies at the frequency of satellite overpass, dramatically enhancing our monitoring capabilities relative to in situ observation networks. To date, however, satellite missions targeting salinity have focused on open ocean rather than coastal applications. NASA's Aquarius mission, launched in June 2011, and the European Space Agency's Soil Moisture and Ocean Salinity (SMOS) mission launched in November 2009, are capable of measuring sea surface salinity from space across the world's oceans, but the 150 km spatial and 7-day temporal resolution of Aquarius and the 250 km spatial and 10-30 day average temporal resolution of SMOS is too coarse for coastal and estuarine environments. The Chesapeake Bay, for example has a maximum width of only 48 km (NASA, 2011). The coarse resolution of these salinity missions stands in contrast to the 1km spatial resolution estimates of sea surface temperature (SST) that are produced with near global coverage on a daily basis by MODIS and other sensors. Estimating high-resolution coastal and estuarine surface salinity from satellite is known to be a valuable tool, yet no proven or operational salinity algorithm exists for the Chesapeake Bay.

Attempts to successfully map sea surface salinity via remote sensing have ranged from Skylab photography (Lerner et al., 1977) to microwave radiometer measurements (Blume et al., 1978), decametric wave ranges (Kachan et al., 1997), ESTAR measurements, and Landsat TM data (McKeon et al., 1976). The use of satellite imagery to map sea surface

salinity in an estuary was first performed in the San Francisco Bay by Khorram et al. (1982). This pioneer study found correlations between Landsat TM color bands and sea surface salinity in an estuarine environment. Other studies (Bowers et al., 2008; Maisonet et al., 2009) explore the empirical relationships between colored dissolved organic matter (CDOM) and salinity using remotely sensed ocean color in a coastal setting. These studies showed that a straight-line relationship between CDOM and salinity is expected dependent on the ratio of the flushing time of an estuary and the timescale of the source variation.

The empirical relationship between colored dissolved organic matter (CDOM) and salinity is important in that CDOM serves as an intermediary function between remote sensing reflectance bands and sea surface salinity. This relationship assumes that fresh-high CDOM river waters mix conservatively with salty-low CDOM seawater, and therefore an inversely correlated relationship between CDOM and salinity. Since we can measure CDOM from space, we can also derive salinity values from remotely sensed observations. It is important to note that this method only works in systems in which there is conservative mixing between coastal waters and rivers. Flocculation and photodegradation could invalidate the assumptions of conservative mixing in this method, however previous work (Del Castillo et al., 1999; De Souza Sierra et al., 1994) has shown that these effects have negligible impacts on CDOM concentration. Therefore, because sea surface salinity is a function of colored dissolved organic matter, it is also a function of remote sensing reflectance. Thus we are confident in our assumption that sea surface salinity can be expressed directly as a function of remotely

sensed ocean color bands. To minimize the number of empirical models applied when deriving salinity from satellite registered radiance, and to capture any additional information on salinity contained in MODIS reflectance bands, we used the standard remote sensing reflectance bands in a multivariate regression model rather than a univariate model using solely CDOM.

The purpose of this study is to predict sea surface salinity in the Chesapeake Bay at 1km resolution using MODIS-Aqua ocean color bands (Table 1.2). This effort is built on work by Geiger et al. (2012), in which Chesapeake Bay salinity fields were estimated at 1km resolution using an artificial neural network (ANN) algorithm applied to MODIS-Aqua data. Here, we test the hypothesis that salinity predictions with smaller or similar errors can be achieved using simpler, more transparent statistical models. To explore a range of statistical modeling options, this study uses eight empirical models typically used when representing continuous response variable data. The eight statistical models are: a Categorical and Regression Tree model (CART), a Generalized Linear Model (GLM), a Generalized Additive Model (GAM), a Random Forest Model, a Mean model, an Artificial Neural Network (ANN), a Multivariate Adaptive Regression Spline (MARS), and a Bayesian Additive Regression Tree (BART). Each of these models includes the dependent response variable[3] sea surface salinity and a set of remotely sensed independent predictor variables[4] described in the data description section below. To test the generalizability of model-predicted sea surface salinity across the diverse

[3] In a statistical experiment, a "dependent response variable" is the observed variable whose changes are determined by the presence of one of more independent variables (Brownlee, 1960)

[4] A independent predictor variable" is a manipulated variable who presence determines the change in the dependent variable (Brownlee, 1960).

salinity conditions of the Chesapeake Bay, we run six seasonal and regional cross validation tests using the top three performing salinity models. The spatial and temporal cross evaluation lends to a more generalizable salinity product than earlier Chesapeake Bay salinity products.

2.2. Data Description

2.2.1. Study Area

The Chesapeake Bay is the largest estuary in the United States, extending 332 km (from Havre de Grace, MD to Cape Charles, VA) along the Atlantic Coast (Figure 2.1). The Chesapeake Bay estuary has a strong north-to-south salinity gradient that includes oligohaline (0-6[5]), mesohaline (6-18), and polyhaline (18-30) zones (Baird et al., 1989). Sea surface temperatures in the Bay range from local wintertime lows of -0.5°C to summertime highs of 31°C. The oligohaline upper Bay has a mean depth of 4.5m, the mesohaline middle Bay 10m, and the polyhaline lower Bay 9m, giving the overall Bay an average depth of 6.5m (22ft) (Baird et al., 1989).

[5] *In situ* and estimated salinity values reported in this study use a standard unitless measure.

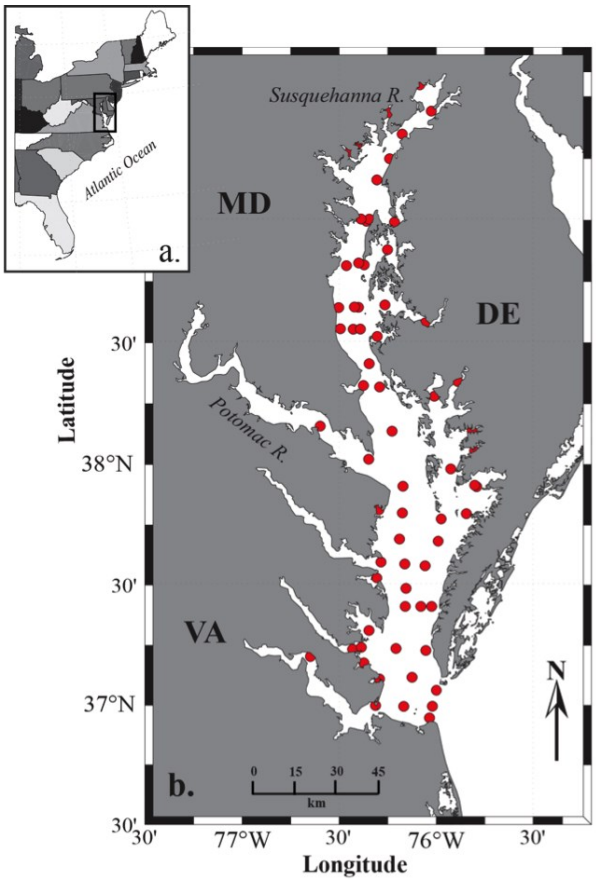

Figure 2.1 The a) Mid-Atlantic coast and the b) Inset of the Chesapeake Bay Estuary with 67 Chesapeake Bay Monitoring Program stations.

The physical transport regime of the Chesapeake Bay estuary follows the classical estuarine circulation model of partially mixed estuaries, in that it is characterized by a 2-layer gravitational circulation scheme. As salt water enters the mouth of the Bay along the eastern shore, there is a net up-estuary flow of water, which occurs below the pycnocline, and a complementary net down-estuary flow as the fresh surface water makes its way from the head to the mouth of the Chesapeake Bay (Pritchard, 1952).

The drainage area of the Chesapeake Bay watershed encompasses 166,000 km^2. Freshwater flows into the Chesapeake Bay estuary from 25 main rivers and tributaries.

The Susquehanna River is the largest tributary in the Chesapeake Bay and accounts for approximately 45% of freshwater flow into the Bay (Baird et al., 1989).

2.2.2. In situ measurements

The analysis performed in this paper made use of in situ environmental data collected by the Chesapeake Bay Monitoring Program (Table 1.1). Bi-monthly data was collected during various research cruises organized by the Maryland Department of Natural Resources (MDDNR) and the Virginia Department of Environmental Quality (VADEQ). The dataset included in situ salinity measurements from 67 monitoring stations (Figure 1.1) along the Bay's axis collected from 2003 through 2010. Using the satellite diffuse attenuation coefficient for down-welling irradiance at 490nm, we calculated the optical depth at each sampling location and found that the mean optical depth of our samples was 0.89m. Therefore, sampling measurements more than 1m in depth were excluded from this study for reasons of remotely sensed surface optical depth.

2.2.3. MODIS satellite measurements

The satellite remotely sensed ocean color products used in this study were from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) Aqua (Table 1.1 & 1.2). Standard ocean color data products were downloaded from NASA's ocean color website (http://oceancolor.gsfc.nasa.gov/), and then batch processed in the SeaWIFS Data Analysis System (SeaDAS). Level-2 daytime standard suite ocean color products at 1km spatial resolution were mapped directly to a cylindrical coordinate system and then

standard quality control flags were applied. Daily satellite images were acquired for the same time period as in situ measurements.

For the purposes of in situ-satellite calibration, we matched in situ station data to the daily satellite measurements within a 1 km radius of the sampling station. Any remotely sensed measurements that were within 1 km of the monitoring station were averaged and thus representative of the unique value of that salinity "pixel". This sampling procedure yielded 620 satellite and in situ matched measurements for use in statistical analysis.

2.3. Methods

2.3.1. Statistical Models

This study presented eight different statistical models developed to predict sea surface salinity via remotely sensed ocean color measurements in the Chesapeake Bay. We chose the eight major types of empirical models that are typically used to regress continuous response variable data. A holdout cross validation was used with the eight statistical models in which 80 percent of the matchup data points was used to train the models and the remaining 20 percent was used for validation. Table 1.2 summarizes the twelve predictor variables that were used to train the eight empirical models presented below. Multivariate models were also compared to a univariate model that used the standard MODIS-Aqua CDOM product (Morel et al., 2009) to predict salinity. The univariate model was found to underperform multivariate models, and will not be discussed further. All statistical computations were carried out in R Statistical Package 2.14 (R, 2011), on an Intel Xeon W3580 Processor, 3.33 GHz machine with 12GB RAM. Computational

time for all statistical models within the holdout validation test was less than one hour, with the exception of the BART model which required up to seven hours of computational time.

2.3.1.1. Generalized Linear Model (GLM)

Generalized linear models are an extension of the standard Ordinary Least Squares (OLS) linear model that allows for regression analysis of both continuous and count data (Nelder et al., 1972). An OLS standard model works to minimize the sum of vertical distances between the observed and predicted response, commonly called the sum of squared residuals (Hastie et al., 1998). An OLS model is composed of two key elements: 1) the random component, which is the probability distribution of the response variable, y, given the predictor variables x_i and 2), the linear predictor, which is an equation that incorporates the data from the predictor variables. A generalized linear model generalizes the standard OLS model by adding a link function, which relates the linear predictor to a function of the predictor variables specifying the conditional mean (Cameron et al., 1998). The link function transforms the expectation of the linear predictor. The salinity measurements in this dataset exhibited a normal Gaussian distribution and therefore we used a normal identity link function $\mu = X\beta$ in the construction of the GLM.

2.3.1.2. Generalized Additive Model (GAM)

A GAM is a flexible statistical model that extends the traditional linear model by allowing for nonlinear relationship between the dependent response and independent predictor variables (Hastie et al., 1986). This model replaces the $X\beta$ link function of the

generalized linear model with a non-parametric smoothing function f(X). The smoothing function can provide information about the relationship between the predictor variables and response variable that is not revealed using a traditional linear model. Nonlinear effects of the covariates on the response variable y can be expressed using GAM. For this study the standard smoothing approach, a cubic regression spline, was used. A cubic regression spline imposes a smoothness on the function f(X), with a potential knot point at each of the unique values of x. Again, an identity link function was used to establish a relationship between the mean value of the response variable y and the smoothed function of the x together with a Gaussian conditional distribution (Hastie et al., 1986).

2.3.1.3. Artificial Neural Network (ANN)

An artificial neural network (ANN) is commonly defined as a massive interconnected network composed of processors, which operate in parallel and learn from experience and training (Lee et al., 1992). The idea of a neural network comes from the biological neural system; the processing elements of an ANN serve as the neurons, while the connections are like synapses from a biological system. The neurons in the ANN are interconnected by means of various information channels. A neural network has at least three basic layers: the inputs, the hidden layer, and the outputs. Input neurons send data via synapses or connections to the hidden layer then via more connections send data to the output neurons (Figure 2.2). Each synapse has an unknown parameter called the "weight"; the weighted inputs are added together and if the sum exceeds the pre-specified threshold then the neuron fires, giving an output (Lee et al., 1992). To maximize prediction accuracy, we first tested two different neural networks, one with 40 hidden nodes and one

with 45 hidden nodes, with these sizes selected based on Geiger et al. (2012). The neural network that exhibited the optimum node size for salinity prediction was then used in the holdout cross validation. In training our ANN models we did note a dependence on the randomly selected initiation points for the weights (i.e., the final trained network varied slightly for different initiation sets)[6]. As a result, we trained 5 different ANNs of each size, with each network starting from a different, fixed seed numbers for the initial sampling of the weights. We report the error from the average of these five models below.

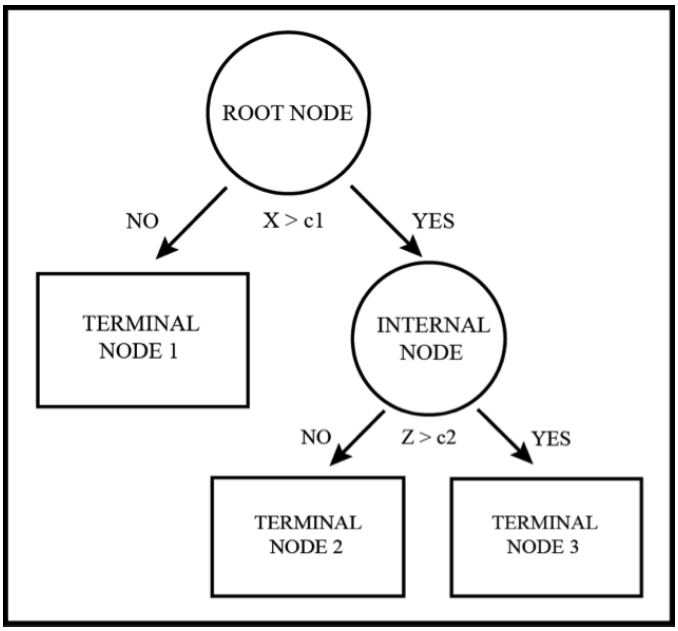

Figure 2.2 Classification and Regression Tree Structure (adapted from Berk et al., 2006).

2.3.1.4. Multivariate Adaptive Regression Spline (MARS)

Multivariate adaptive regression spline (MARS) is a non-parametric regression method that can be seen as an extension of a linear model allowing for interactions and non-

[6] This variability persisted well beyond the number of replications of the ANN training algorithms at which convergence was reported.

linearities in a dataset (Freidman et al. 1991). MARS behaves like a generalized linear model, but based on automatically selected basis functions. MARS builds models in the same fashion as recursive partitioning trees, but allows for a forward and backward pass (Hastie et al., 2008).

2.3.1.5. Tree-Based Data Mining Techniques

To further a different class of models for empirically predicting sea surface salinity in the Chesapeake Bay, the study used four tree-based data mining methods: classification and regression tree (CART) (Breiman et al., 1998), Bayesian additive regression trees (BART) (Chipman et al., 2010), bagged categorical and regression trees (BCART) (Sutton, 2005), and Random Forest model (RF) (Breiman et al., 1998). Each of the four tree-based data mining methods explores the relationship between the predictor variables and the dependent response variable, sea surface salinity. The dataset undergoes recursive binary partitioning at the nodes. Tree-based methods give a flexible description of relationships within the dataset while also providing a convenient visual for result interpretation.

2.3.1.6. Mean Model

Each of the statistical models outlined in this section were compared to a mean statistical null model. Our mean model was simply the average value of the response variable salinity. For validation purposes, all nine models including the mean model were input into the holdout run.

2.3.1.7. Geographic Model

Surface salinity, optical depth, and CDOM/salinity relationships are highly variable and dependent on location in the Chesapeake Bay. To test the added value of using ocean color bands, as well as the correlation between salinity and geographic location, a holdout validation test using only latitude and longitude was run employing the nine statistical methods outlined above.

2.3.2. Cross-validation of top statistical models

In order to develop the most effective remotely sensed salinity prediction model, we needed to test the generalizability of the empirical algorithms in the Chesapeake Bay. To validate the reliability of the salinity predictions throughout the Chesapeake Bay, we split up the in situ-satellite matchup dataset temporally and geographically. In both cases, the top three statistical models, determined by lowest mean absolute errors (MAE), were used in a cross-validation on different spatial and temporal periods of the Bay.

There is great seasonal and geographic variability in in situ salinity, fresh water discharge (Figure 2.3), as well as in cloud cover that interferes with satellite retrieval of surface reflectances in the Bay. Therefore, although we developed and tested the statistical models on year-round in situ-satellite matchups, it is important to train the models on one season and predict for another to reflect the variations in fresh water inflow into the Chesapeake Bay. To do so, we divided the entire matchup dataset into two discharge datasets: high (December through May) and low (June through November). As described above, there are various salinity regimes throughout the Bay,

which exhibit certain characteristics dependent on geographic location and biophysical processes in that location. For example, cold saline seawater is characteristic of water lying along the mid-eastern shore due to estuarine circulation patterns in the Chesapeake Bay. To cross-check this spatial variability, we also split in situ-satellite matched data points spatially, into North versus South and East versus West. Geographic divisions were performed separately from seasonal divisions. For both, the top three statistical models were trained on one database (low/South/East) and then tested on the other (high/North/West), then vice-versa.

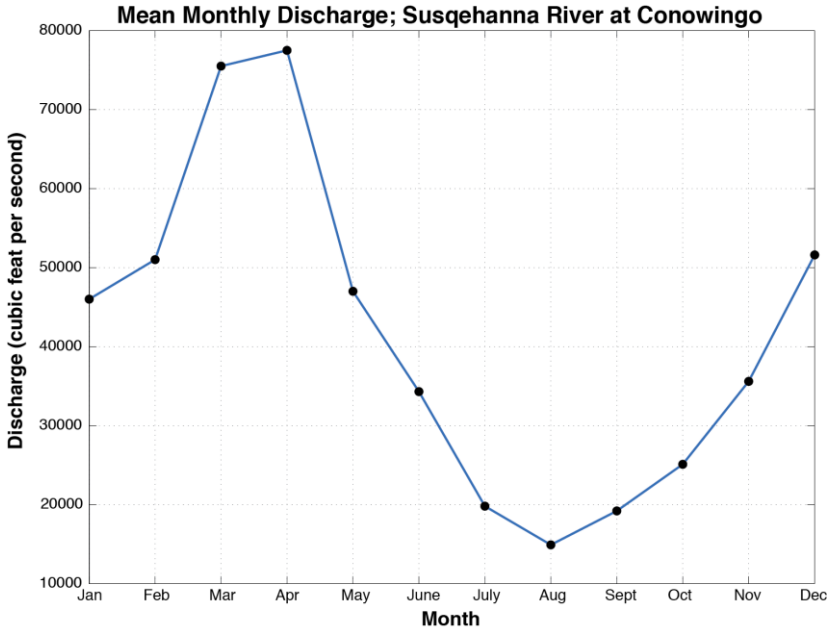

Figure 2.3 Mean Monthly (1970-2000) Discharge at USGS 01578310 Susquehanna River at Conowingo, MD Station (adapted from USGS, 2012).

2.3.3. One-to-one comparison of remotely sensed versus in situ salinity

To assess the functionality of our empirical salinity model, we tested the top-performing model on a separate set of remotely sensed independent variables. To guarantee a one-to-

one comparison with in situ salinity measurements, we chose a daily MODIS image with good spatial coverage from a day (September 18, 2006) when the Chesapeake Bay Monitoring Program conducted in situ salinity measurements. Overlap in MODIS and station measurements from that day allowed for 13 in situ-satellite comparison points.

2.4. Results and Discussion

2.4.1. Model Comparisons

The in situ-satellite dataset was fit with the eight statistical models outlined above using a repeated holdout validation test. Each of the statistical models were compared to the mean prediction model in the holdout test to determine how well each model performed assuming the dataset mean salinity value. This results in 36 pair-wise tests with a mark of statistical significance if the p-value[7] on a given test is less than 0.00014 in accordance with the needed Bonferroni correction (Devore, 1995). As shown in Table 1.3, all eight statistical models outperform the mean model by a statistically significant amount ($p < 2.2e-16$). The generalized additive model has the best prediction accuracy with the lowest MAE of 1.82 followed by the 45-node ANN model with a MAE of 1.85, and the GLM with a MAE of 1.93. The one to one regressions of the matched in situ salinity vs. the model predicted salinity for the GAM, and the ANN models (Figure 2.4) show that there are approximately ten data points in which the prediction model clearly under predicts the true salinity value. The locality of these data accounts for the large error as it was found that each outlying predictor was nearby the mouth of a fresh water tributary. Not only do we see increased freshwater flow, but also variability in the discharge of

[7] P-value indicates the probability that the result obtained in a statistical test is due to chance rather than a truce relationship between measures (Brownlee, 1960).

sediments, terrigenous organic matter, detritus, and chlorophyll concentrations in these regions. These changes can complicate the bio-optical properties of the water due to the absorptive properties of CDOM, phytoplankton mass, and detritus, which further affect the shape of the remote sensing signal at each location. Further model development and variable specification needs to be carried out to understand the effects of these environmental conditions on model prediction.

GAM, followed by ANN, also has the highest predictive accuracy when judged by mean square error (MSE) and root mean square error (RMSE) (Table 1.4). The difference in MSE values between GAM and ANN is not significant at a 95% confidence level. All empirical models outperform the mean model with respect to MSE. MSE and RMSE are useful metrics for identifying outliers in the model fit. A RMS error of equal or higher value than the MAE (see Table 1.5) indicates that there are outlier salinity outputs in the top three salinity prediction models. It is important to note that there is no statistically significant difference in the salinity prediction for the GAM, the ANN, or the GLM (see Table 1.6).

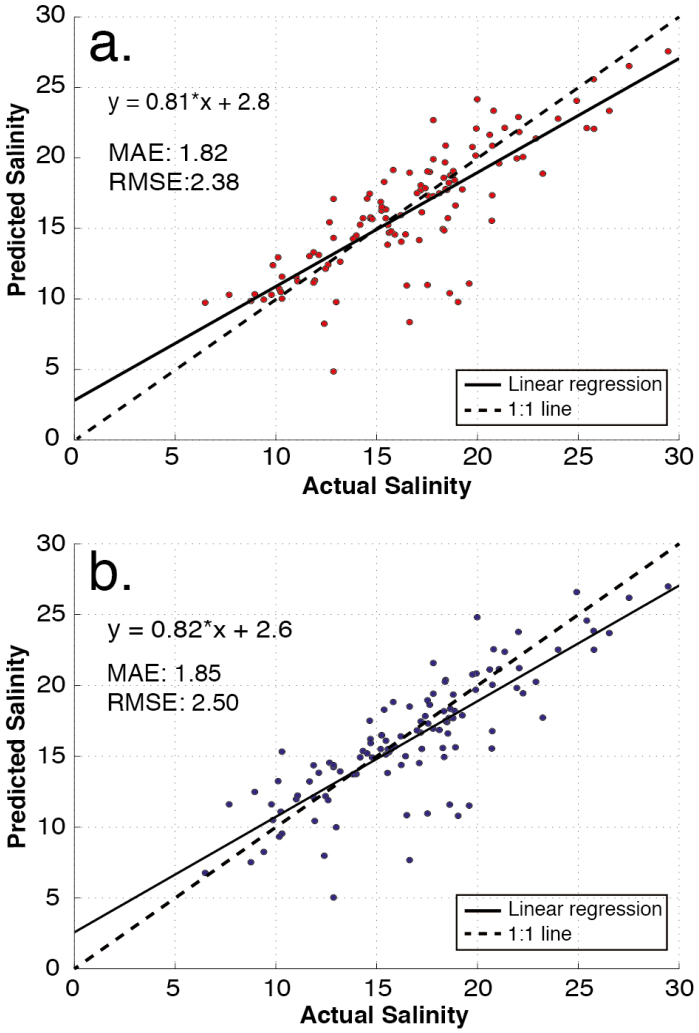

Figure 2.4 One-to-one model regression between in situ salinity and predicted salinity for a) the GAM and b) the ANN. The mean absolute errors for each statistical model are: a) 1.83 and b) 1.85

For GAM, it is also possible to examine the specific importance and influence of each of the reflectance bands in the prediction of salinity. Table 1.7 lists the p-values associated with each smoothed term in the GAM. Nine of 12 variables included in the GAM are statistically significant (p-value 0.05). Though model results show that latitude and longitude are the most significant predictor variables in the GAM model, a holdout run

using only latitude and longitude shows a significant (p-value <0.05) decrease in prediction accuracy (Table 1.8), and thus value added in using the remotely sensed reflectance values. While all but three of the predictor variables are statistically significant and thus important in predicting the response variable y, not all of the variables that have high importance are highly influential to the model outcome[8]. Of the twelve smoothed terms included in the GAM, half show high influential behavior on the predicted response. Indicative examples of variable responses are shown in Figure 2.5. Remote sensing reflectance (Rrs) at 488nm is positively associated with salinity (Figure 2.5a), while Rrs at 667nm and at 443nm is negatively associated (Figure 2.5b, 2.5c). Other predictor variables, such as Rrs at 412nm, are statistically significant in the GAM but show no particularly strong independent influence on salinity (Figure 2.5d).

[8] Variable reduction was performed on both the GAM and GLM, but this did not improve the prediction accuracy over the final non-reduced models.

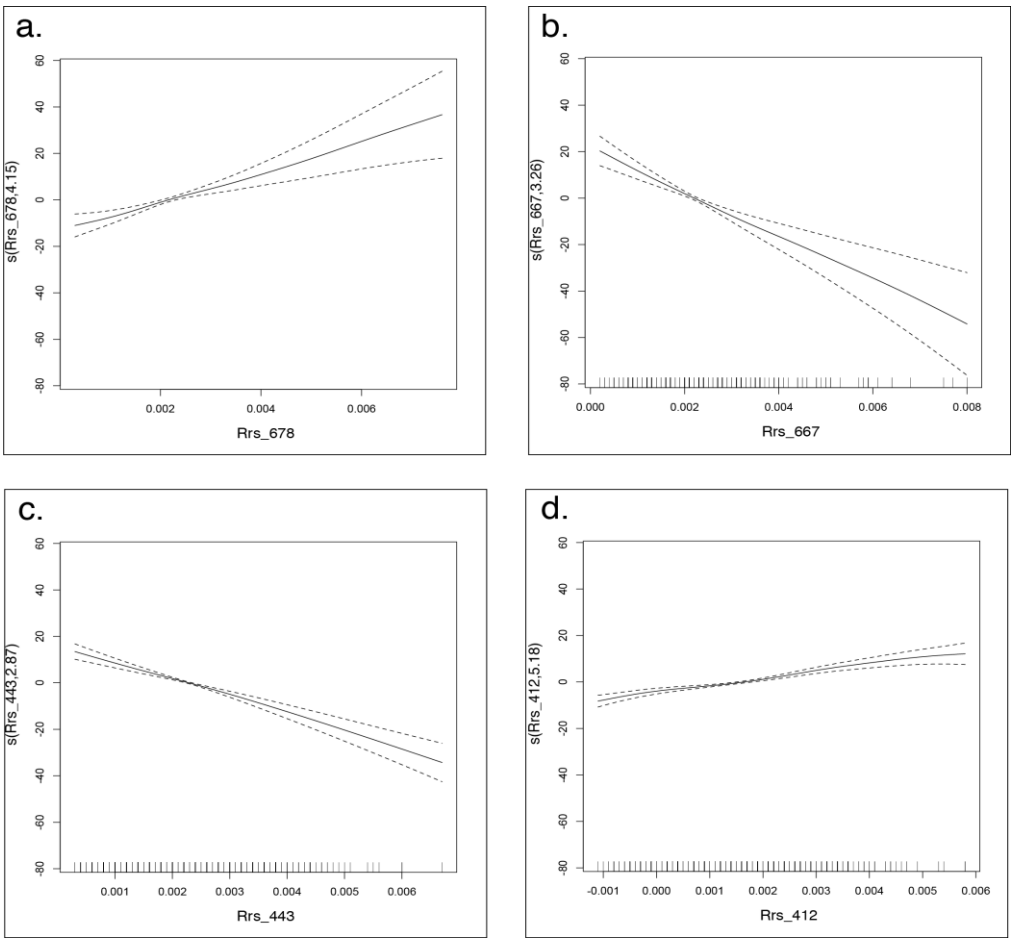

Figure 2.5 Variable Plots for GAM model; a) Rrs_678, b) Rrs_667, c) Rrs_443, and d) Rrs_412

2.4.2. Cross-validation of models

To test the generalizability of our remotely sensed salinity product in the Chesapeake Bay, we ran six seasonal and regional cross validation tests using the top three performing salinity models. In these cross-validation analyses, the GAM and GLM perform with better error accuracy than the ANN in all cases but one (Table 1.9). The first two of the six cross-validation tests evaluated the generalizability of salinity models from east to west in the Bay. In training the three models on the eastern Bay portion and testing on the West and vice-versa, GAM performs the best with a MAE of 1.8 in the first

case, and GLM the best when trained on the West and tested on the East (note that differences between GAM and GLM were not statistically significant). When the same tests were conducted for low and high, all three models perform well—when trained on high for low testing, both GLM and GAM has a MAE of 2.3. While the generalizability of the models for East versus West and low versus high perform well in terms of low MAE and RMSE values, the cross-validation tests for North versus South are not as consistent in their prediction results. From Table 1.9, we can see that although the GAM MAE for "North for South" performs equally as well as the previous tests, the model trained on the South and tested on the North underperforms relative to the mean model. This is the only generalizability test for which either GAM or GLM was outperformed by the mean model. This result is likely a product of systematic differences between the relatively fresh North and the saltier South, and is the subject of continued investigation.

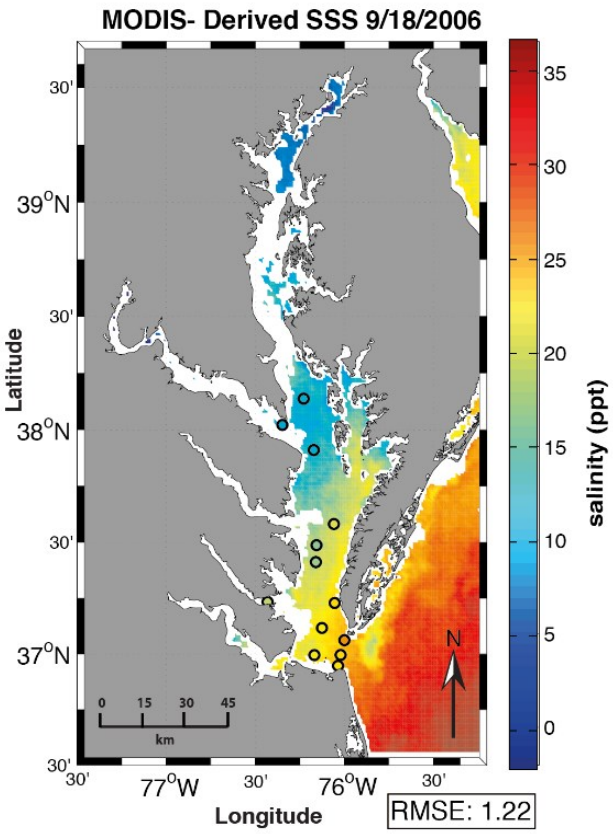

Figure 2.6 GAM predicted salinity for September 18, 2006 with in situ station locations and actual salinity values marked by colored in black circles.

2.4.3. One-to-one daily GAM predicted, in situ comparison

The comparison of in situ salinity to GAM predicted salinity for September 18, 2006 results in improved prediction accuracy over the holdout validation data sets. Five of the 18 in situ stations were removed from the one-to-one comparison because they fell outside remote sensing coverage for the given day. Figure 2.6 shows the predicted GAM salinity for the entire Bay, as well as the actual in situ salinity at the stations marked by filled circles. The RMS error improved from 2.38 in the holdout validation tests to 1.22 for the daily prediction versus in situ. Figure 2.7 shows the regression of the in situ versus GAM predicted salinity with a slope of 0.89. In addition to the improved RMS

error between actual and predicted salinity, predicted salinity from the GAM follows a believable salinity regime for the Bay. Not only do the predicted values fall within the natural range for the Bay, but also the prediction actually exhibits the spatial gradients explained earlier in this paper.

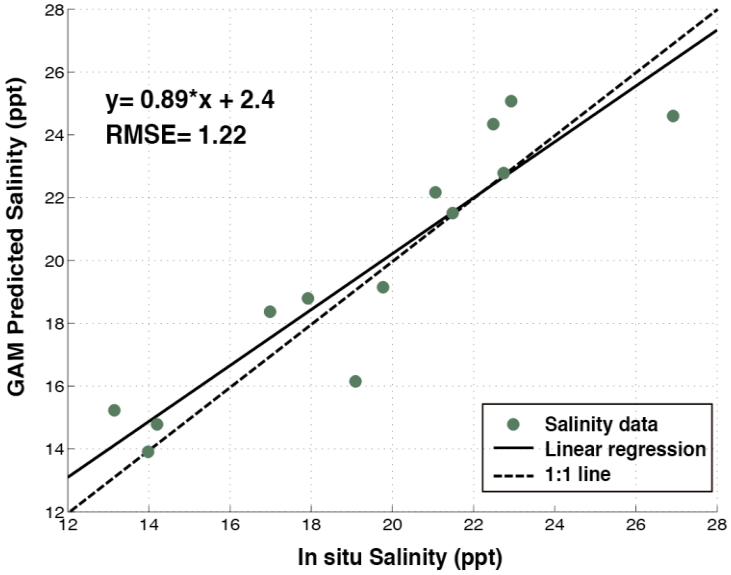

Figure 2.7 Regression between in situ salinity and GAM prediction salinity for September 18, 2006. The RMSE is 1.22.

2.5. Conclusions

The eight statistical models presented above show that remotely sensed products can be used to accurately estimate sea surface salinity in the Chesapeake Bay. While predicting salinity via remote sensing for the Bay is still in its beginning stages, the results of applying these models to remotely sensed measurements can provide the imperative missing block to many biological and physical marine applications. Three models that perform particularly well in estimating salinity were the generalized additive model, the generalized linear model, and the artificial neural network.

Additionally, six cross-validation tests were run to evaluate the generalizability of our salinity estimates across various temporal and spatial regimes in the Chesapeake Bay. Table 1.8 summarizes the MAE and RSME results from the six cross-validation models. From the prediction results we can conclude that for the Chesapeake Bay, the GAM and GLM outperform the artificial neural network; further supporting our original hypothesis that a more transparent model can estimate sea surface salinity with equal or better accuracy than an ANN. We can assume that the tendency of the more complicated neural network was to over fit the data, resulting in the poor prediction accuracy, showing that the transparent models like the GLM and GAM are more generalizable to the Chesapeake Bay region.

The empirical models presented in this study are particularly good at estimating sea surface salinity in the Chesapeake Bay. We do note, however, that salinity estimates were found to be highly dependent on geographic location. Results show that latitude and longitude are the most significant predictor variables in the nine surface salinity estimation models. While this locality issue was anticipated for the Chesapeake Bay and thus accounted for, it indicates that attention to mixing processes, fresh water inflow, and seasonality will be required when applying these statistical salinity models to other coastal regions. A second limitation of the study is in the data itself. The in situ salinity measurements presented in the paper was taken at a water depth of approximately 0.5m. Satellite remote sensing is useful in detecting sea surface reflectance signals, but the inability to penetrate below the ocean's surface and clouds often limits the availability of

data. Therefore lies a discrepancy between the depth of the in situ measurement and the remotely sensed surface reflectance. Further work will focus on interpolation methods to understand salinity changes as a function of water column depth. A third limitation of the model's training data is the temporal and spatial scarcity of in situ salinity measurements. As presumed, the availability of remotely sensed reflectance data far exceeds the number of environmental surface measurements.

In order to obtain full temporal and spatial coverage of Chesapeake Bay, the satellite remote sensing data and in situ observations can be combined with a fluid dynamical model through data assimilation. In this way, the observations are utilized when they are available, but model dynamics will drive accurate forecasts in the absence of observations. Data merging of in situ and RS observations through the use of a numerical model will provide a full 3 dimensional coverage of the Bay that will therefore allow us to propagate the satellite sea surface information deeper into the water column. Such a data assimilation system is being developed for the Chesapeake Bay (Hoffman et al., 2011) and in future work we hope to leverage that system and create more complete sea surface salinity estimations for the Bay.

Table 1.1 Data types, spatial resolution, temporal resolution, and sources of data
* L2 Modis AQUA standard suite of products (see Table 1.2).
[a] Maryland Department of Natural Resources
[b] Virginia Department of Environmental Quality

Data Type, Parameters (Period of Record)	Spatial Resolution	Temporal Resolution	Source
In situ station data, salinity, surface temperature (2003-2010)	67 main-stem monitoring stations on Bay axis, 1m vertical resolution	~20 surveys/yr, bi-monthly to monthly cruises	MDDNR[a]; VA DEQ[b] (Chesapeake Bay Monitoring Program)
L3-mapped ocean color and thermal SST satellite products* (2003-2010)	1 km spatial resolution	Daily satellite overpasses	Modis AQUA, National Aeronautical Space Administration

Table 1.2 Variables used in model development

Variable Name	Variable Description	Mean (μ)	Standard Deviation	Maximum	Minimum
Salinity (predictor variable)	*In situ* salinity measurement at surface	16.49	4.69	31.65	0.00
Lat	Latitudinal data coordinate of *in situ*-satellite matchup	37.68	0.51	39.44	37.00
Lon	Longitudinal data coordinate of *in situ*-satellite matchup	-76.14	0.15	-75.79	-76.46
Rrs_412	Remote sensing reflectance at 412-nm	0.0014	0.0012	0.0058	-0.001
Rrs_443	Remote sensing reflectance at 443-nm	0.0022	0.0011	0.0067	0.0003
Rrs_469	Remote sensing reflectance at 469-nm	0.0029	0.0013	0.0083	0.0006
Rrs_488	Remote sensing reflectance at 488-nm	0.0035	0.0015	0.0094	0.0008
Rrs_531	Remote sensing reflectance at 531-nm	0.0055	0.0020	0.0126	0.0018
Rrs_547	Remote sensing reflectance at 547-nm	0.0060	0.0021	0.0140	0.0018
Rrs_555	Remote sensing reflectance at 555-nm	0.0059	0.0020	0.0139	0.0019
Rrs_645	Remote sensing reflectance at 645-nm	0.0030	0.0015	0.0145	0.0006

Rrs_667	Remote sensing reflectance at 667-nm	0.0022	0.003	0.0137	0.0002
Rrs_678	Remote sensing reflectance at 678-nm	0.0022	0.0012	0.0135	0.0003

Table 1.3 Comparison of Holdout Mean Absolute Errors (MAEs) Based on 120 Random Holdout Samples

* P-values in **bold** represent statistically significant differences between models

Model	MAE	p-value: GAM	p-value: CART	p-value: BCART	p-value: RF	p-value: MEAN	p-value: ANN	p-value: BART	p-value: MARS
GLM	1.93	3.4e-06	**2.2e-16**	**2.2e-16**	**1.5e-05**	**2.2e-16**	0.0006	**0.0001**	0.1407
GAM	1.82		**2.2e-16**	**2.2e-16**	**4.8e-15**	**2.2e-16**	0.2575	**5.9e-14**	**2.3e-09**
CART	2.39			0.7254	**2.2e-16**	**2.2e-16**	**2.2e-16**	**2.2e-16**	**2.2e-16**
BCART	2.38				**2.2e-16**	**2.2e-16**	**2.2e-16**	**2.2e-16**	**2.2e-16**
RF	2.06					**2.2e-16**	**7.7e-12**	0.5489	0.0015
MEAN	3.72						**2.2e-16**	**2.2e-16**	**2.2e-16**
ANN	1.85							**9.5e-11**	**2.1e-06**
BART	2.04								0.0093
MARS	1.98								

Table 1.4 Comparison of Holdout Mean Squared Errors (MSEs) Based on 120 Random Holdout Samples

* P-values in **bold** represent statistically significant differences between models

Model	MSE	p-value:	p-value:	p-value:	p-value:	p-value:	p-value:	p-value:	p-value:

		GAM	CART	BCART	RF	MEAN	ANN	BART	MARS
GLM	6.40	0.0002	**2.2e-16**	**2.2e-16**	0.0003	**2.2e-16**	0.8135	0.0415	0.7253
GAM	5.67		**2.2e-16**	**2.2e-16**	**1.6e-10**	**2.2e-16**	0.1956	**5.3e-08**	0.0004
CART	9.17			0.6968	**9.7e-15**	**2.2e-16**	**1.1e-07**	**2.2e-16**	**2.2e-16**
BCART	9.08				**9.7e-14**	**2.2e-16**	**2.6e-07**	**2.2e-16**	**2.2e-16**
RF	7.14					**2.2e-16**	0.0800	0.0606	**6.9e-05**
MEAN	22.07						**2.2e-16**	**2.2e-16**	**2.2e-16**
ANN	6.28							0.3101	0.9162
BART	6.77								0.0137
MARS	6.33								

Table 1.5 Comparison of Holdout MAEs, RMSEs, and MSEs

	GAM	ANN	GLM	CART	BCART	RF	MEAN	BART	MARS
MAE	1.82	1.85	1.93	2.39	2.38	2.06	3.72	2.04	1.98
RMSE	2.38	2.50	2.53	3.03	3.01	2.67	4.69	2.60	2.52
MSE	5.67	6.28	6.40	9.17	9.08	7.14	22.07	6.77	6.33

Table 1.6 Comparison of Mean Predicted Salinity Based on 120 Random Holdout Samples

	Mean Salinity	p-value: GLM	p-value: GAM	p-value: ANN
In situ	16.73	0.476	0.495	0.394
GLM	16.31		0.986	0.875
GAM	16.32			0.864
ANN	16.22			

Table 1.7 Approximate significance of GAM smoothed terms
*P-values in **bold** represent statistical significance ($p<0.05$)

Smoothed Term	p-Value
Latitude	**2.20e-16**
Longitude	**2.20e-16**
Rrs_678	**4.95e-05**
Rrs_667	**4.27e-08**
Rrs_645	**0.007**
Rrs_555	0.118
Rrs_547	0.293

Rrs_531	**1.16e-06**
Rrs_488	**1.11e-11**
Rrs_469	0.289
Rrs_443	**2.94e-14**
Rrs_412	**3.18e-11**

Table 1.8 LAT-LON only model comparison of holdout MAE and RMSE values.
Values in **bold** represent the models that are significantly different ($p<0.05$) than the original eight models.

	GAM	ANN	GLM	CART	BCART	RF	BART	MARS
MAE	**2.36**	**2.38**	**2.55**	2.41	2.42	**2.40**	**2.36**	2.35
RMSE	**2.98**	**2.98**	**3.21**	3.05	3.05	**3.01**	**2.96**	**2.98**

Table 1.9 MAE and RMSE values for Cross-validation tests
* Naming convention for cross-validation is as follows: "East for West" translates to model trained on East dataset and tested on West dataset.

	MAE				RMSE			
	GLM	GAM	ANN	MEAN	GLM	GAM	ANN	MEAN
East for West	2.1	1.8	3.7	3.3	2.6	2.3	4.7	4.0
West for East	2.6	2.8	4.0	4.1	3.3	3.5	5.2	5.3
North for South	3.4	2.1	5.9	5.7	4.2	2.8	7.0	6.8
South for North	3.0	6.4	6.1	5.7	4.2	9.9	7.1	6.5
Winter for Summer	2.1	2.2	4.0	3.9	2.7	2.8	5.2	4.8
Summer for Winter	2.2	2.0	3.6	3.5	2.9	2.7	4.5	4.6

3. CHAPTER 3: GEOSPATIAL INTERPOLATION OF MODIS-DERIVED SALINITY AND TEMPERATURE IN THE CHESAPEAKE BAY[9]

ABSTRACT

In dynamic coastal systems such as the Chesapeake Bay, limited coverage and frequency of in situ measurements often makes generalizability of regional forecasting systems difficult. Satellite-derived environmental variables have the potential to address this problem, but satellite datasets suffer from incomplete coverage as well: atmospheric conditions—most notably cloud cover—lead to data gaps that significantly hinder the broad application of satellite-informed predictions. In this study, the Chesapeake Bay estuary was used as a model "test bed" to which we applied the power of near real-time satellite-derived observations to the issue of water quality monitoring. To use remote sensing in support of spatially complete estimates of salinity and temperature in the Bay, we tested geospatial interpolation techniques as a method for filling gaps and minimizing errors in the satellite record. These interpolated values were then compared to the output of a regional hydrodynamic model in order to assess the relative value of each method for generating inputs into various modeling applications. Results show that MODIS-derived salinity and temperature can be interpolated with acceptable accuracy in the Bay, with a mean absolute error of 1.88 ppt and 0.60 °C. These errors differed systematically from ChesROMS errors both spatially and seasonally, with higher errors for salinity and lower errors for temperature at most sampling stations throughout the year. This suggests that

[9] Urquhart E., Hoffman M., Murphy R., Zaitchik B. (2013) Geospatial Interpolation of MODIS-Derived Salinity and Temperature in the Chesapeake Bay. *Remote Sensing of Environment.* 135: 167-177.

the two techniques offer complementary information that can be applied to ecological monitoring systems in complex estuaries like Chesapeake Bay.

3.1. Introduction

Estuaries and coastal waters are dynamic environments, subject to variable currents and mixing processes that produce high temporal and spatial variability in water properties relevant to hydrodynamics, water quality, and ecology. Estuarine and coastal environments are also increasingly vulnerable to adverse environmental, biological, and societal change under pressures of human population growth, sea level rise, land degradation, and climate change. In coastal regions such as the Chesapeake Bay, for example, it has been documented that both the abundance and distribution of harmful organisms is increasing throughout various near-shore regions (Marlyand Department of The Environment, 2010).

The highly variable and evolving nature of these environments makes them notoriously difficult to survey and monitor. As conditions continue to change in poorly characterized and unpredicted ways, there is a vital need for more advanced and spatially complete information than can be provided from traditional grab-sample monitoring networks. To meet this need, researchers and environmental managers have employed three approaches that enable improved spatial continuity in coastal systems: interpolation of in situ data, hydrodynamic numerical modeling, and satellite remote sensing.

Spatial interpolation can be used in coastal water bodies that have existing in situ monitoring networks—such as the Chesapeake Bay—to estimate the value of an

environmental parameter at un-sampled locations based on measured values at other locations (Murphy et al., 2010). Various research efforts (Bahner, 2006; Chehata et al., 2007; Hagy et al., 2004; Murphy et al., 2010) have used spatial interpolation methods to analyze water quality trends over space and time in the Chesapeake Bay. These methods can offer critical information for environmental applications like water quality assessment and management, but they are themselves limited by the spatial extent and temporal coverage of the operating in situ network: in the absence of an adequate number of sampling points, no interpolation is possible. Additionally, even when interpolations can be generated, estuaries like the Chesapeake Bay often present complications for spatial interpolation. Complex hydrodynamics create a system in which observations from a limited number of in situ samples may not capture the variety of conditions associated with confined currents, river plumes, and other localized features within the Bay.

These challenges for statistical interpolation are one of the motivations for applying deterministic numerical models such as the Chesapeake Bay Regional Ocean Modeling System (ChesROMS; (Xu et al., 2011)) to environmental monitoring. These models are able to simulate water properties continuously in space and time as a function of physical, chemical and biological processes. Hydrodynamic models have been used for many years and have proven to be beneficial in informing coastal management applications (Committee on Environmental and Natural Resources, 2010). However, due to structural and parameter uncertainties in complex coastal systems it is often difficult to implement

and evaluate such hydrodynamic models. Even after years of development, known and unknown biases can persist in a modeling system that have no obvious solution.

A third common coastal monitoring approach, satellite remote sensing, provides observational data for assessment and analysis of complex coastal systems in the absence of in situ measurements. Satellite remote sensing is typically the most cost-effective and efficient means of data acquisition in regions that are inaccessible, distant, dangerous, or too large for traditional monitoring approaches. Various studies have used remote sensing to address particular hypotheses about coastal environments, including water quality monitoring (Del Castillo & Miller, 2007; Li et al., 2003; Ondrusek et al., 2012), habitat mapping (Ortiz & Tissot, 2008), and oil spill mitigation (Leifer et al., 2012), to name just a few applications. However, satellite data alone does not offer complete coverage on account of cloud cover, coastline interference, and other issues related to data quality and resolution. Figure 3.1a shows the minimum, mean, and maximum percent satellite coverage by month for the Chesapeake Bay for 2003 to 2010, illustrating decreased satellite coverage during the summer months. Furthermore, Figure 3.1b shows the mean monthly percent satellite coverage for 12 in situ stations that span the Bay's north-to-south salinity gradient (see Figure 3.2). Limited coverage in summer is problematic for the Chesapeake Bay, as these months are characterized by increased biological activity, increased hypoxia, and strong vertical stratification—conditions that make reliable monitoring particularly important.

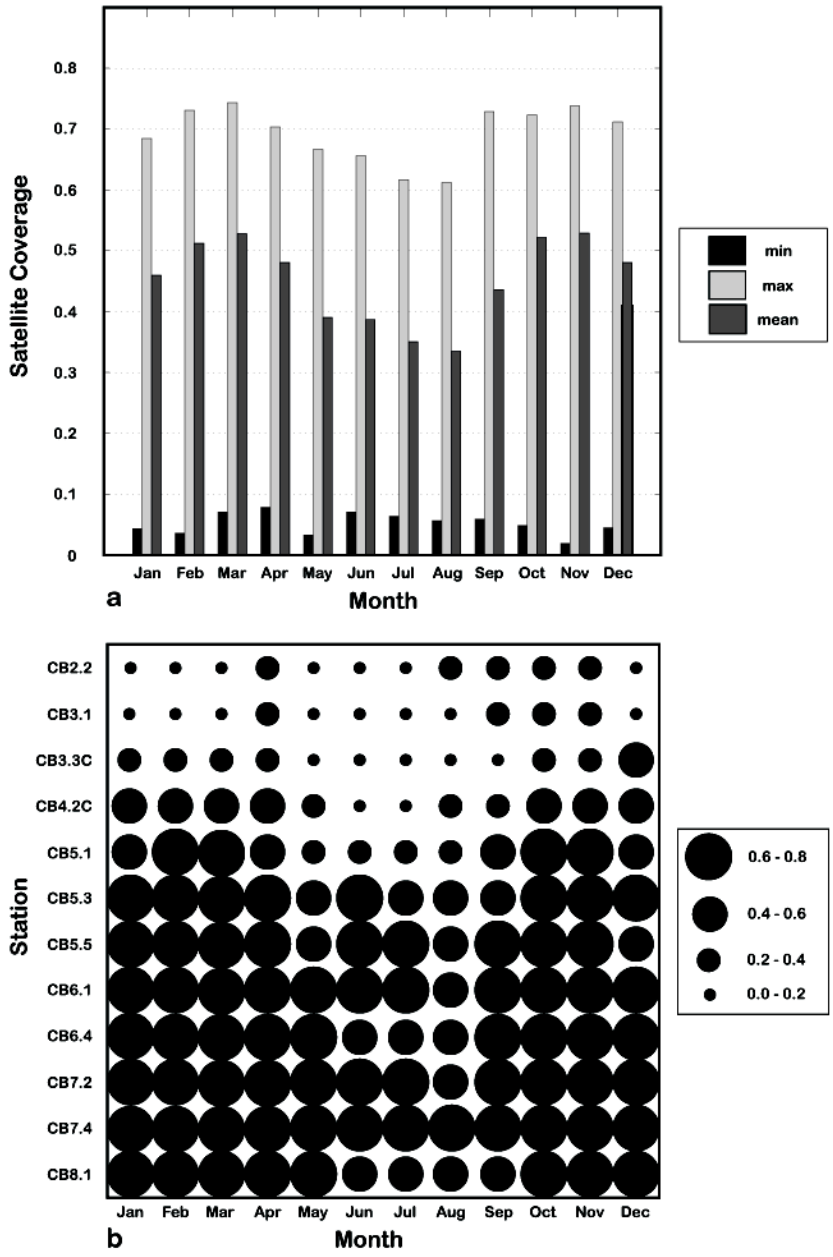

Figure 3.1 a) Minimum, mean, and maximum fraction monthly satellite coverage, and b) fraction monthly mean satellite coverage by station for the Chesapeake Bay, 2003-2010

Recognizing the values and the limitations of these three approaches to spatially complete coastal monitoring, we present analyses designed to merge remote sensing and geospatial

interpolation techniques to generate spatially and temporally complete estimates of water surface temperature and salinity in the Chesapeake Bay. These merged estimates are then compared to similar estimates from the ChesROMS hydrodynamic model in order to identify relative strengths and weaknesses, to consider implications for empirical ecological models currently used in the Bay, and to inform ongoing efforts to optimize satellite data assimilation in ChesROMS. It is important to note that for the purposes of this study, the Chesapeake Bay is used as a model "test bed." Furthermore, given the availability of data, the methods presented here can be applied to the issue of water quality monitoring in other coastal regions.

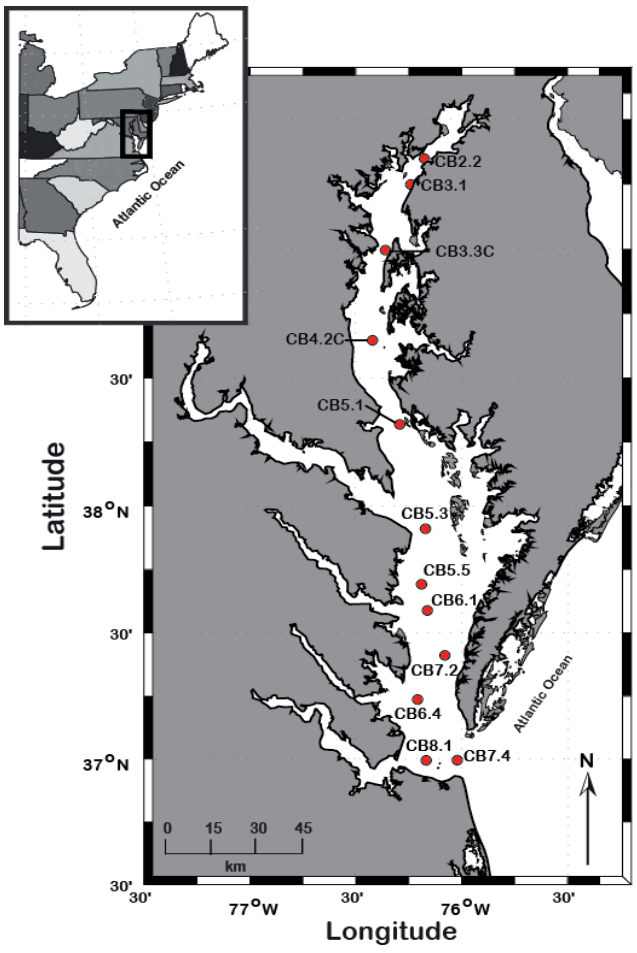

Figure 3.2 Chesapeake Bay and in situ sampling locations used in this study; station CB2.2 corresponds to the upper bay with the lower bay at CB8.1

3.2. Methods

3.2.1. Study area

The Chesapeake Bay is the largest estuary in the United States, extending 332 km (from Havre de Grace, MD to Cape Charles, VA) along the Atlantic Coast (Figure 3.2). The upper Bay has a mean depth of 4.5 m, the middle Bay 10 m, and the lower Bay 9 m, giving the overall Bay an average depth of 6.5 m (22 ft) (Baird & Ulanowicz, 1989). Its width ranges from 5.5 km near Aberdeen, MD to 56 km near the mouth of the Potomac River (Chesapeake Bay Program, 2012b). The Chesapeake Bay estuary has a strong north-to-south salinity gradient that includes oligohaline (0-6 psu), mesohaline (6-18 psu), and polyhaline (18-30 psu) zones. Freshwater flows into the Chesapeake Bay estuary from 25 main rivers and tributaries. The Susquehanna River is the largest tributary in the Chesapeake Bay and accounts for approximately 45% of freshwater flow into the Bay (Baird and Ulanowicz, 1989). Not only do we see increased freshwater flow in these tributary regions, but also variability in the discharge of sediments, terrigenous organic matter, detritus, and chlorophyll concentration. This can affect the bio-optical properties of the water due to the absorptive properties of colored dissolved organic matter (CDOM), phytoplankton mass, and detritus, which further affect the shape of the remote sensing signal in these regions. Sea surface temperatures in the Bay range from local wintertime lows of -0.5°C to summertime highs of 31°C (Baird & Ulanowicz, 1989). The Chesapeake Bay is a partially mixed estuary, characterized by a relatively

large freshwater input with density-driven circulation resulting in a two-layer structure consisting of a fresh seaward-flowing surface layer and a saline return-flow beneath (Xu et al., 2002). This study focuses on the mainstem portion of the Chesapeake Bay. The mainstem Bay was selected to minimize complications relating to coastal interference with satellite retrievals and to interpolation within nonconvex tributaries.

3.2.2. MODIS satellite measurements

The remotely sensed temperature and empirically derived salinity measurements used in this study were from NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) on board the Earth Observing System Aqua satellite. Sea surface temperature (SST) is derived from MODIS thermal infrared (IR) channels in which IR radiances are transformed (through use of the Plank function) to units of temperature (°C). MODIS "skin" temperature measurements are made available in a variety of spatial resolutions and temporal periods (Brown & Minnett, 1999). There was good agreement between the MODIS and in situ temperature measurements with a mean absolute error (MAE) of 0.67°C in the Chesapeake Bay for 2003 to 2010. Sea surface salinity was statistically derived using a Generalized Additive Model (GAM), in which longitude, latitude, and the MODIS ocean color products were used as independent predictor variables. The MODIS-derived salinity product estimated surface salinity with a MAE of 1.82 relative to mean Chesapeake Bay salinity of 16.5. Further details of the described empirical salinity algorithm can be found in Chapter 2. It is important to note the underlying difference between remotely sensed temperature, which is derived from MODIS thermal infrared satellite measurements, and satellite-derived salinity, which is statistically

derived from geographic coordinates and MODIS ocean color products for the Chesapeake Bay. Daily satellite images were acquired for the same time period as the bi-monthly in situ measurements available from the Chesapeake Bay Program (Chesapeake Bay Program, 2012a). Standard ocean color and temperature products were obtained from NASA's ocean color website (http://oceancolor.gsfc.nasa.gov/), and batch processed using the SeaWIFS Data Analysis System (SeaDAS). For the purposes of this study, daily Level 2 daytime standard suite ocean color products at 1-km spatial resolution were mapped directly to a cylindrical coordinate system and then standard quality control flags were applied. For cross-validation purposes and interpolation comparisons, satellite observations that fell within two days of the in situ sampling measurements were included in our analysis. This sampling procedure yielded 1040 salinity and 855 temperature matches between satellite and in situ measurements for use in statistical analysis.

3.2.3. In situ measurements

The analyses performed in this paper made use of in situ environmental data collected by the Chesapeake Bay Monitoring Program (Chesapeake Bay Program, 2012a). Bi-monthly data was collected during research cruises organized by the Maryland Department of Natural Resources (MDDNR), the Virginia Department of Environmental Quality (VADEQ), and various universities. Salinity and water temperature are both measured at fixed station locations using an YSI probe (see CBP Water Quality Data Dictionary for more probe information, (Chesapeake Bay Program, 1993)). The dataset used in this study included in situ measurements from 12 monitoring stations (Figure 3.2)

along the Bay's axis collected from 2003 through 2010. These 12 stations were selected based on their geographic location and temporal record over the seven-year study period. These 12 stations span the Bay's north-to-south salinity gradient, with four stations in upper-middle Bay, four stations in the middle Bay, and four stations near the mouth of the Chesapeake Bay (Figure 3.2). Using the satellite diffuse attenuation coefficient for down-welling irradiance at 490nm (Kd_490), we calculated the optical depth at each sampling location and found that the mean optical depth of our samples was 0.89m. Therefore, in situ sampling measurements more than 1m in depth were excluded from this study because they are deeper than the remotely sensed surface optical depth. For the purposes of this paper, the in situ water quality data described above was used to validate interpolated satellite fields.

3.2.4. Chesapeake Bay Regional Ocean Modeling System (ChesROMS)

ChesROMS is a freely available, open source model that is used in numerous research and monitoring applications (Constantin de Magny et al., 2009; Prasad et al., 2011; Xu et al., 2012). The configuration used in this study is identical to that used in Hoffman et al. (2012). An outline of the model configuration is presented below, but for a more complete description of ChesROMS see Xu et al. (2012) and Hoffman et al. (2012). The model grid is a 100x150 curvilinear horizontal mesh (Figure 3.3), which corresponds to a resolution of 1-5 km with finer resolution in the upper Bay. There are 20 vertical sigma levels, with finer resolution near the surface and bottom. The vertical resolution changes with depth, which goes from approximately 2 to 35 m. The top modeled layer is taken to be the ChesROMS surface measurement. The depth of this measurement is typically around 0.05 m and is therefore closer to surface (and satellite measurements) than many of the in situ measurements.

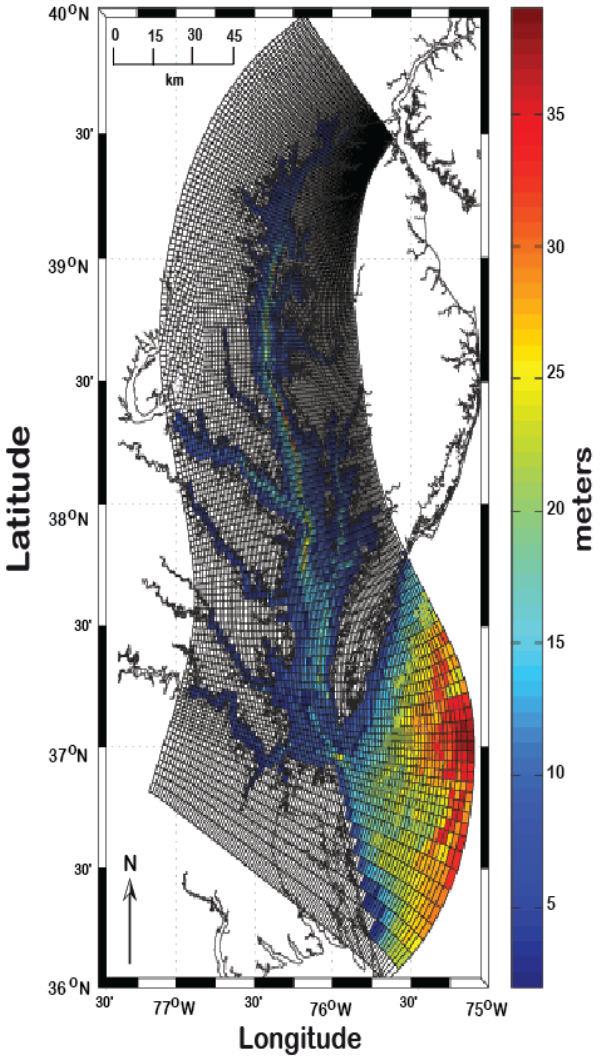

Figure 3.3 ChesROMS grid and Bathymetry; adapted from Hoffman et al. (2012)

Bathymetry data for ChesROMS is drawn from the US Coastal Relief Model at National Oceanic and Atmospheric Administration's National Geophysical Data Center (NGDC) (Bahner, 2006). At the open ocean boundary, tidal water level is set using nine tidal constituents from the Advanced Circulation Model (ADCIRC) EC2001 tidal database

(Mukai et al., 2002) and non-tidal water levels interpolated from stations in the NOAA National Ocean Service program. River forcing is prescribed daily for nine tributaries based on the United States Geological Survey (USGS) stream water monitoring project. Finally, 3-hourly winds, net shortwave and downward longwave radiation, temperature, relative humidity, and pressure at the air surface boundary are prescribed by the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR).

For this study, water quality and hydrodynamic predictive output from ChesROMS was obtained for the years 2003 and 2007. The selected output contains estimates of surface temperature and salinity every six hours for each model grid cell. Cells containing one of the 12 in situ monitoring stations were identified and the model output at 12 pm EST (time closest to satellite overpass) was extracted so that the in situ station data, interpolated data, and ChesROMS data used in cross-comparisons would share spatial and temporal characteristics to the greatest extent possible. Modeled temperature and salinity were evaluated against in situ data and then cross-compared to interpolated values from satellite-derived temperature and salinity for the same study period.

3.2.5. Spatial Interpolation Methods

The satellite-derived salinity and temperature data in the Chesapeake Bay were spatially interpolated using two versions of the geostatistical kriging method: ordinary kriging (OK) and universal kriging (UK). Kriging is a method for generating statistically optimal

spatial predictions (Cressie, 1993). The estimates from kriging are weighted averages of the observations:

$$\hat{Y}(s_0) = \sum_{i=1}^{n} w(s_i) y(s_i) \quad (1.1)$$

where $\hat{Y}(s_0)$ is the interpolated value at location s_0, $y(s_1)...y(s_n)$ are observed values at locations s_1 to s_n, and $w(s_1)...w(s_n)$ are the weights which are generated from a model of the spatial correlation structure of the data, typically one of several valid variogram models that is fit using the observations. More details on kriging and variogram fitting can be found in numerous texts (Cressie, 1993; Diggle & Ribeiro, 2007; Schabenberger & Gotway, 2004), plus in a previous study on which this kriging implementation was based (Murphy et al., 2010)

Relevant to this study, the kriging method can be formulated as a general linear regression model (Diggle & Ribeiro, 2007)

$$Y(s) = \beta_0 + \beta_1 X_1(s) + \cdots + \beta_p X_p(s) + \varepsilon(s) \quad (1.2)$$

with s representing location, $Y(s)$ the parameter being interpolated at s, $X_1...X_p$ potential covariates indexed by location, $\beta_1...\beta_p$ fitted coefficients, and $\varepsilon(s)$ the random error that is assumed to have a multivariate correlation structure that can be represented by the variogram. When one or more covariates are included in Eq. 2, it is referred to as universal kriging. If no covariates are included (i.e., $Y(s) = \beta_0 + \varepsilon(s)$), it is referred to as ordinary kriging and all of the variation in the estimates comes from the correlation structure represented in $\varepsilon(s)$. Note that historically the term "universal kriging" referred to kriging with the coordinates (e.g., latitude and longitude) as covariates, but in what is called the "model-based" approach, any spatially-varying auxiliary information can be

included as a covariate (see discussion in Murphy et al. (2010)). All universal kriging interpolations performed in this study included NGDC bathymetry and latitude covariates at each grid cell.

3.2.6. Spatial Error Analysis

To find the dominant spatial error patterns in our interpolated data, we applied Empirical Orthogonal Function (EOF) analysis to the difference between interpolated and in situ temperature and salinity (i.e., the "error" in interpolated satellite data relative to in situ observation). EOF analysis defines a linear combination of uncorrelated variables, chosen to capture the maximum observed variance contained in the original, possibly correlated data. That is, given multiple observations of a data matrix x, EOF analysis finds vectors that are linear combinations of the elements of xs (Jolliffe, 2002). For a more detailed description of EOF analysis, the reader is referred to Jolliffe (2002), Lorenz (1956), and Wilks (2006).

EOF analysis was performed on the error for both satellite salinity and temperature at the 12 sampling station for both OK and UK interpolated parameter values. The time period covered by the error datasets spanned the same period as the original interpolation datasets. We calculated the monthly mean deviation from the in situ parameter value for each sampling station over the duration of the sampling period. This procedure generated a 91X12 array of differences for salinity and 96X12 array of temperature differences for which a scalar product with itself formed two covariance matrices. The matrices were diagonalized and the resulting eigenvalues, referred to as EOF modes, were ordered from

largest to smallest. While EOF modes are simply mathematical, the leading modes can often be associated with underlying dynamical features of a dataset (Preisendorfer, 1988).

3.2.7. Performance Evaluation

One common method used to evaluate the performance of geospatial interpolation is holdout cross validation. Due to the clumped pixel nature of satellite imagery and thus the close proximity of the data locations, holdout techniques were not employed. Instead interpolated satellite data was compared to in situ observations using several "error metrics" in order to evaluate interpolation performance. Mean error (ME), mean absolute error (MAE), and root mean squared error (RMSE) are three possible difference measures that are used to identify outliers in model fit, as well as to evaluate each interpolator's performance. It is important to note that in this study we interpolated satellite output, but used in situ data to evaluate the interpolation's accuracy. Therefore, it is possible that there are two sources of errors: (1) satellite-data mismatch due to timing of the match or the satellite estimate itself (see Figure 3.1), and (2) error in the interpolation method alone. In the case of the universal kriging, these errors can be moderated by the value of added information in the independent covariates.

3.2.8. Computational Details

All interpolation computations were carried out in the statistical package R 2.14 (R Development Core Team, 2008) using the geoR contributed package (Diggle & Ribeiro, 2001) on an Intel Xeon W3580 Processor, 3.33 GHz machine with 12 GB RAM. Computational time for yearly interpolations (average of 44 files per year) was less than

three hours. Mapped salinity and temperature satellite images within two days of in situ sampling, which also met a minimum pixel threshold of 600 non-missing pixels within the Bay, were imported into the R environment. After importing the satellite data into R, images were mapped to the ChesROMS 100x150 curvilinear grid cells. For the universal kriging method, both the latitude and total water column depth at each location were used as independent model covariates. Both latitude and depth, independent of each other, are known to be related to temperature and salinity in the Bay. In general, the salinity in the Bay follows a decreasing northward gradient that decreases with depth. For temperature, depending on the season, shallow waters are typically warmer than deeper waters (Chesapeake Bay Program, 2012b). For both ordinary and universal kriging, a variogram model was estimated for each satellite pass.

The large volume of data included in this study demanded an automated method to estimate variograms for kriging. The automated method (similar to Murphy et al. (2010)) fit each dataset to both an exponential and spherical variogram model using inferences based on restricted maximum likelihood (Cressie, 1993). The best fitting model (spherical or exponential) was selected for each satellite dataset based on the restricted log-likelihood values. For both the exponential and spherical models, the range and sill values were estimated using a matrix of initial values to begin the maximization process. The range of initial values for the variogram sill was between 0.5 and 1.5 of the original variance of the data. Initial values for the variogram range were between 10 and 200% of the maximum distance between data cells. The initial value for the variogram nugget was fixed at zero.

All performance evaluation, EOF analysis, and model comparisons were carried out in the MATLAB computing environment (The MathWorks Inc., 2010). For the EOF analysis, singular value decomposition (SVD) (Linz & Wang, 2003) was used to find principal components.

3.3. Results

3.3.1. Variograms

The results in Table 2.1 show that no one variogram model best fits all of the datasets. While the linear model was not one of the pre-specified variogram models, it is found that both the exponential and spherical models may produce a linear fit in presence of very large sill and range values. For that reason, a variogram was labeled linear when the practical range distance was greater than 1.5 times the maximum distance between samples (Murphy et al., 2010). Summary statistics of the different variogram models revealed that variograms fit using both the OK and UK methods for salinity most commonly followed a linear shape. For water temperature, 56 percent of the OK variograms exhibited a linear trend, while a spherical model was most common in the UK variograms. Results show that for both salinity and temperature the OK method had more data sets with the linear model and fewer datasets with the exponential and spherical models, indicating that both parameters display spatial correlation that decreases with increased distance. Similarly, though a linear model is the dominant UK model type for salinity, there is a more even distribution between the three model types for both parameters.

3.3.2. Comparison of Spatial Interpolation Methods

For a comparison of the two interpolation methods, both sets of validation results were analyzed to find trends in each interpolator's performance. Table 2.2 shows the average performance metrics for OK and UK evaluated against in situ observations. For both temperature and salinity, all performance measures indicate that on average the UK method outperformed the OK method. This result was expected, as the universal kriging method allows for the addition of covariates upon which the environmental parameters depend. The satellite-derived salinity model described in Chapter 2 is known to slightly over-predict surface salinity in the upper regions of the Chesapeake Bay. Importantly, results from this study show that the UK MAE of 1.88 is consistent with the MAE found for the original statistically derived salinity product.

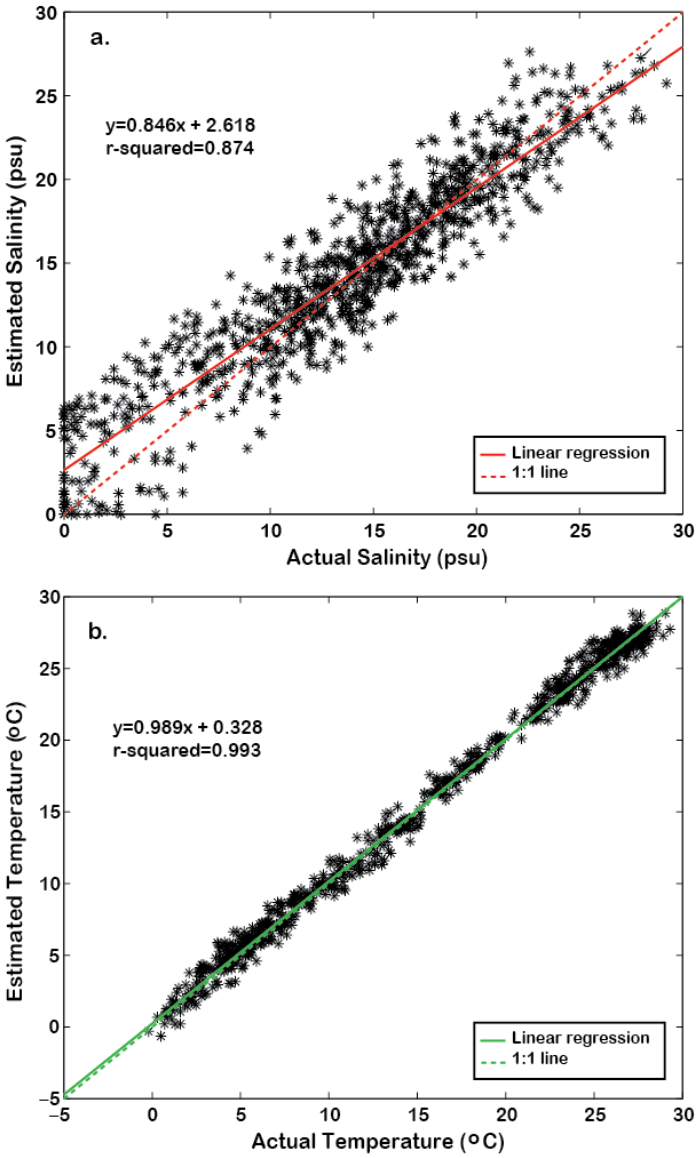

Figure 3.4 One-to-one model regression between actual and estimated a) salinity, and b) temperature; solid line shows linear regression line; dashed line one-to-one line

The interpolation error measures presented in Table 2.2 are averages over the entire study period. Figure 3.4, which shows in situ versus UK estimated salinity and temperature for each station and comparison event, provides further insight on error characteristics. The comparison between interpolated salinity and in situ salinity (Figure 3.4a) reveals that at

low salinity values the correlation decreases and the interpolated salinity over-predicts the in situ value. These low salinity values all come from sites near the head of the Bay or close to tidal tributaries. These results are consistent with the analyses in Chapter 2 of MODIS-derived salinity, which showed that the algorithm used in this study is prone to error in the upper Chesapeake Bay. This suggests that errors found in the interpolation output here are likely due to the remote sensing product rather than the interpolation method. Validation errors averaged by station for OK, UK, and the non-interpolated MODIS-derived salinity estimate (RS) (Figure 3.5a), confirm this pattern, as the largest errors are found in the upper region of the Bay, particularly the two northernmost stations. For water temperature (Figure 3.5b), both OK and UK interpolators showed decreased performance at the upper and lower Bay stations. This decrease in performance is likely due to limited satellite observations in the upper Bay. When implementing kriging, we assume a constant mean value for the parameter and model all variability as functions around this constant mean (Murphy et al., 2010). When using UK, this constant mean assumption is replaced by a linear trend. Therefore, in the absence of upper Bay satellite data, the interpolated estimates using the UK method should be used cautiously because they are extrapolations based on this linear trend from observations in the middle of the Bay. Using the OK method, estimates beyond the boundaries of the satellite data will tend toward the mean of the observed values. The validation errors averaged by month (Figure 3.5c,d) show the extent to which seasonality also plays a role in interpolation performance. For surface salinity, the OK interpolator suffered from large errors in summer, while the UK interpolator's performance was fairly constant throughout the year. For water temperature, both OK and UK interpolators had

increased performance during the fall and early winter months. Improved performance in these months is not surprising, since satellite data coverage is most extensive during these less-cloudy months (Figure 3.1). It is notable, however, that UK performance on salinity does not vastly degrade during summer months despite limited data coverage. During these data-limited months, the addition of covariate information in UK is particularly valuable.

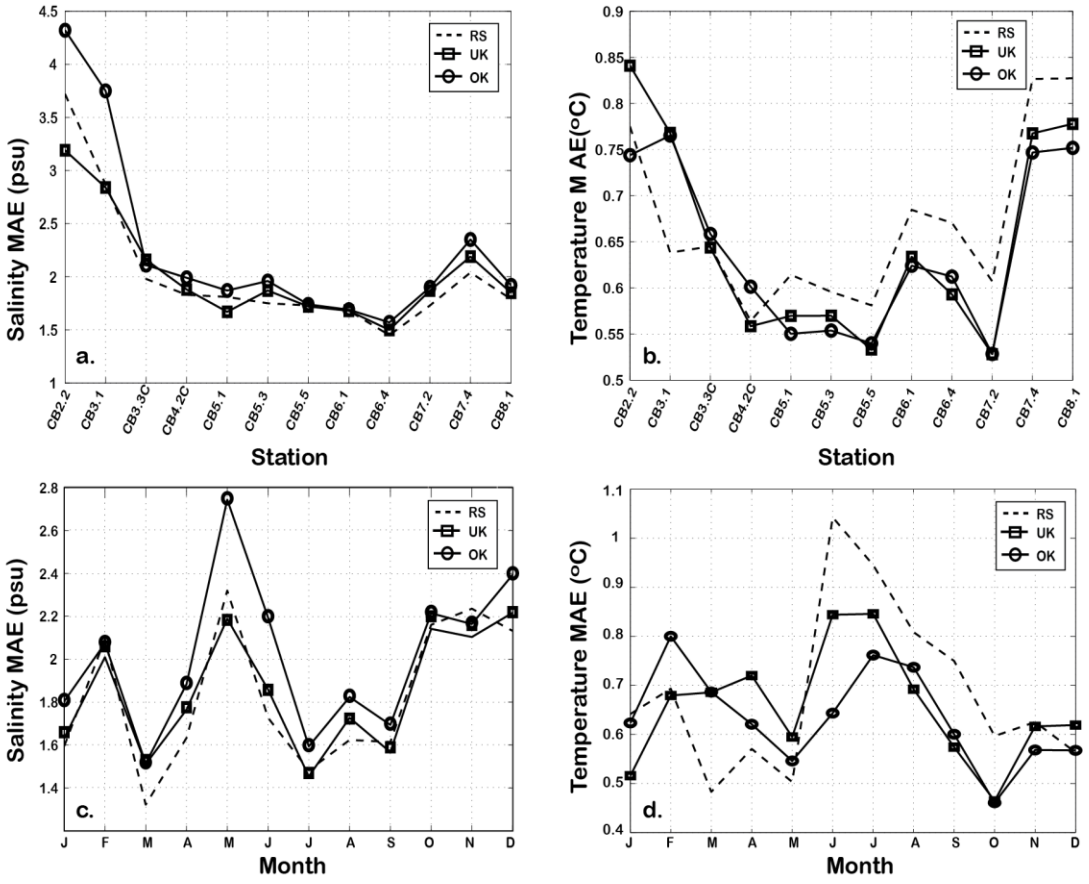

Figure 3.5 Average MAE for ordinary (OK) and universal kriging (UK) and non-interpolated remote sensing (RS) methods by a & b) station, and c & d) month

3.3.3. Spatial Variability of the Leading EOF Modes

For all EOF analyses, results show that the first two EOF modes account for more than 90% of the total error variance in OK and UK salinity and temperature estimates at in situ sampling stations (Table 2.3). Figures 6a and b show the spatial patterns of the first two EOF modes and their corresponding coefficients. The spatial structure of EOF1 for salinity (Figure 3.6a), explaining 79% and 69% of the variance for OK and UK respectively, represents variability that is in phase across the entire Bay but with much higher amplitude in the upper Bay. The second EOF, explaining 13% and 21% of the total variance for OK and UK, shows a north-south error dipole with a maximum in the upper Bay that is out-of-phase with leveled off variation in the remainder of the Bay. While EOF1 and EOF2 have similar spatial patterns for the OK and UK methods, OK has a higher percent variance explained by EOF1, which can be attributed to decreased interpolation performance in the upper Bay region. The spatial pattern of EOF1 for temperature (Figure 3.6b), explaining 66% and 77% of the variance for OK and UK, mirrors that for salinity, with whole-bay variability being amplified in the upper bay. The pattern for EOF2, explaining 24% and 17% of the total variance, is dominated by strong upper Bay and lower Bay regions of the opposite sign. A temporal analysis of the leading salinity EOF modes showed no discernable seasonal trend across the Chesapeake Bay. There was however, a slight peak in variance during summer months for the top OK and UK modes. For temperature, the dominant spatial mode parallels that of salinity with in-phase error variability across the Bay with a high in the upper Bay (Figure 3.6b). The temporal analysis of the temperature modes revealed a more apparent seasonal trend with a peak from late summer into the fall.

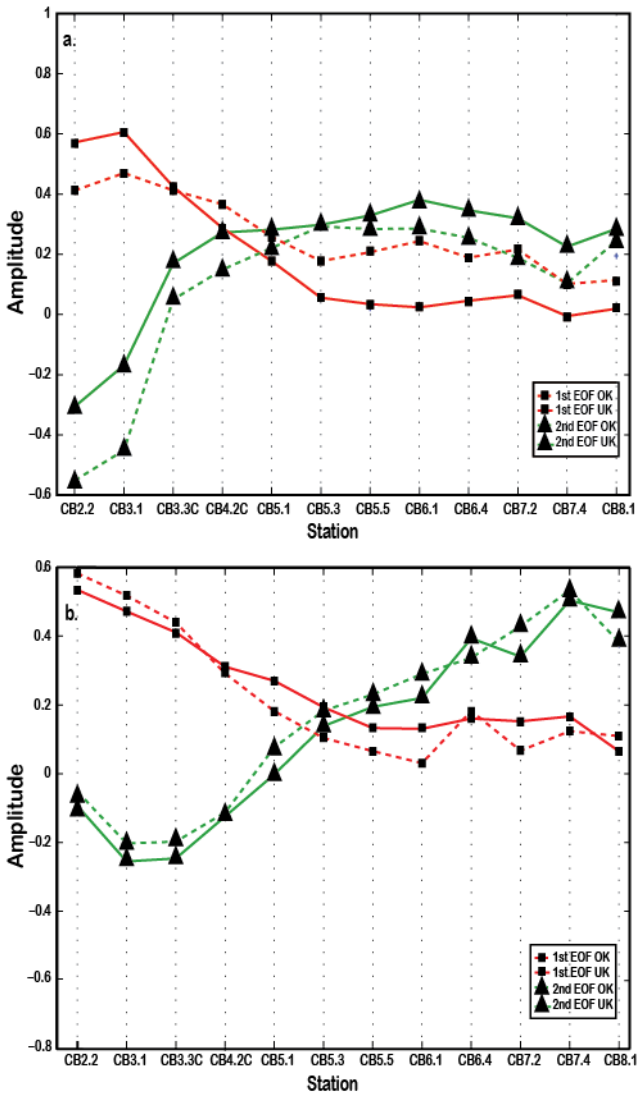

Figure 3.6 Empirical Orthogonal Functions for a) salinity, and b) temperature, for ordinary and universal kriging methods by station; filled in square represents EOF1, filled in triangle represents EOF2, solid line represents OK, dashed line represents UK

3.3.4. Model Comparison

MODIS interpolations were compared to ChesROMS output for 2003 and 2007, the two years for which we had access to ChesROMS model runs. Table 2.4 shows the average error metrics from validation tests for all of the estimation methods. For salinity, all error

measures indicate that ChesROMS performs better, on average, than the kriging methods. The mean error (ME) for interpolated salinity reveals that interpolation techniques tend to overestimate surface salinity, while ChesROMS underestimates salinity throughout the Chesapeake Bay (Table 2.4). For temperature, all error metrics show that both of the kriging methods outperform ChesROMS.

To explore these results further, the average performance errors were summarized by station and month. Figure 3.7 shows the average mean absolute errors (MAE) for UK and ChesROMS temperature and salinity for each station and each month, averaged across the two comparison years (2003 and 2007). For surface salinity, results show poor interpolation performance in the upper Bay and in summer months, with mixed performance in the remainder of the Bay (Figure 3.7a). ChesROMS has on average, lower errors than UK at all stations during all months, with the exception of September (this is likely due to poor ChesROMS performance following Hurricane Isabel in September 2003) (Figure 3.7b,c,d). For water temperature, results show mixed interpolation performance throughout the Bay, with poor performance in the middle and upper Bay during summer months (Figure 3.7e). These results were not unexpected, as the satellite temperature product on which the interpolation was performed is known to have relatively large errors during data-sparse months typically characterized by increased cloud cover. ChesROMS has larger absolute error at most evaluation times and locations, with exceptions in the upper and lower Bay during summer months (Figure 3.7f,g,h). These results are consistent with Figure 2.5b,d in which we see relatively large interpolation errors in the upper and lower Bay and in summer.

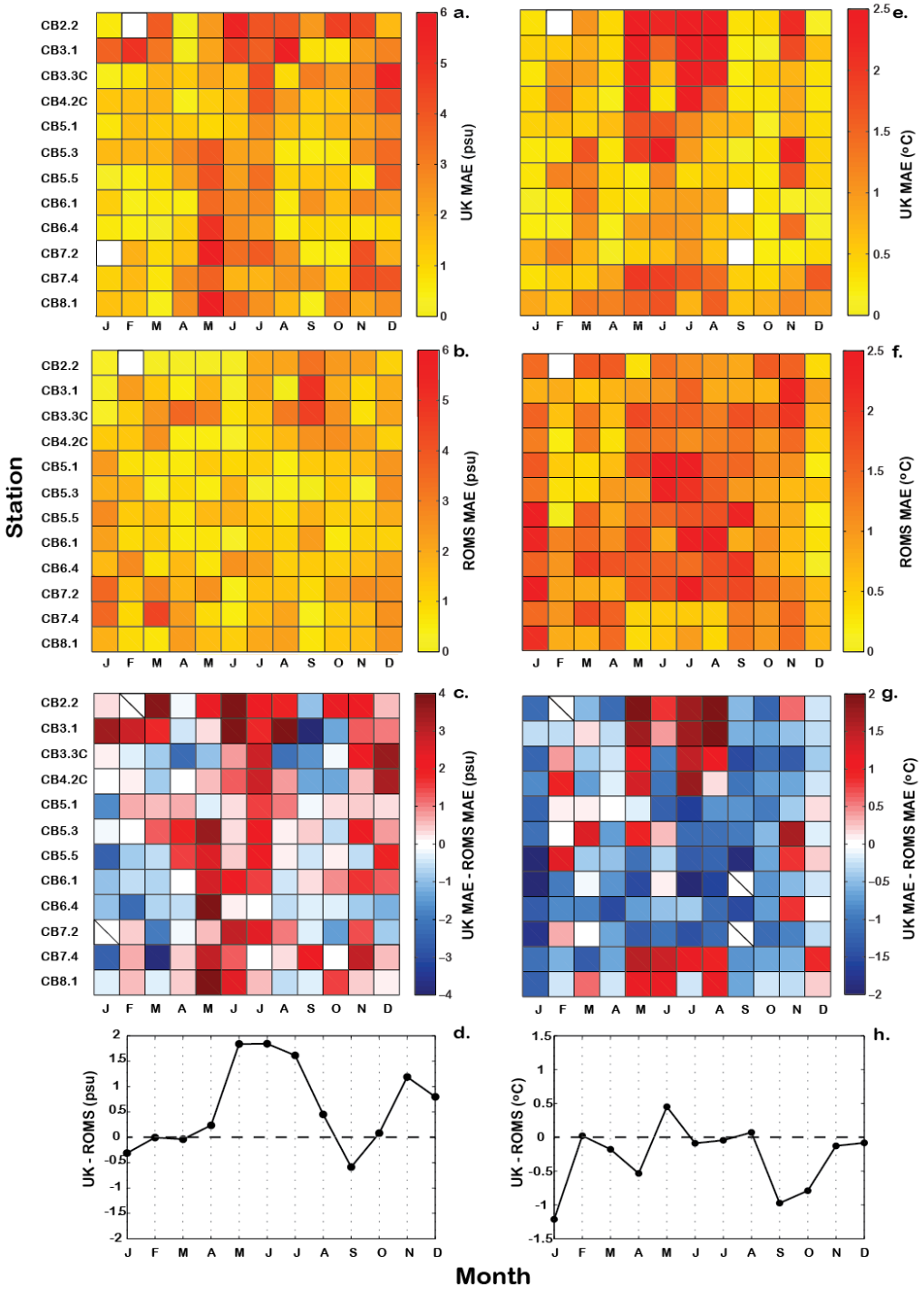

Figure 3.7 Contour plots of MAE for a) UK salinity, b) ChesROMS salinity, e) UK temperature, f) ChesROMS temperature, with MAE difference plots c & d) for salinity

and g & h) temperature by station and month**.** Cells with diagonal lines through them represent missing data

3.4. Discussion and Conclusions

A primary motivation for this study is the fact that satellite observations are spatially and temporally incomplete, resulting in data gaps that limit the utility of satellite-based environmental monitoring. Geostatistical interpolation methods have the potential to overcome this limitation. Here we have compared two methods—ordinary kriging (OK) and universal kriging (UK)—to interpolate satellite-derived temperature and salinity estimates for Chesapeake Bay. Results show that when averaged over seven years of available data, universal kriging outperforms ordinary kriging for both parameters. The change in UK variogram shape (Table 2.1) and interpolation performance (Table 2.2) indicates that including information on latitude and bathymetry does contribute to interpolation performance. Improvements in UK relative to OK were greatest in the upper Bay for salinity (Figure 3.5a) and during summer months for both temperature and salinity (Figure 3.5c,d). Indeed, for salinity, the error of the UK interpolated product in the upper Bay was, on the average, smaller than the error of the non-interpolated remotely sensed (RS) estimate even at locations for which direct satellite observations were available. This indicates that the UK covariates provide significant information for salinity estimates in this hydrodynamically complex portion of the Bay. For temperature, and for salinity everywhere but the very upper portions of the Bay, errors were approximately equal for the interpolated and non-interpolated satellite product, indicating that errors in the interpolated product are primarily inherited from errors in satellite-based

temperature and MODIS-derived salinity estimates; the interpolation techniques did not appreciably add to the total error of the estimate. While depth and latitude were informative covariates in Chesapeake Bay, applications of UK to other coastal water bodies would have to take locally relevant environmental gradients and processes into account—for example, distance from a coastline or major delta.

An EOF analysis of salinity and temperature errors relative to in situ observations identified two dominant spatial EOF modes that accounted for over 90% of the total error variability in interpolated datasets (Table 2.3). The leading salinity mode for both OK and UK exhibited in-phase variability throughout the Bay. The salinity pattern of EOF1 parallels the results displayed in Figure 3.4a and Figure 3.5a, which show poor interpolation performance in the fresh, upper Bay. A time series analysis of the leading EOF modes showed no discernable seasonal trend with a slight peak during late summer for salinity, and a clear seasonal trend with a late summer high for temperature.

A second objective of this study was to perform a comparison of interpolated satellite data and ChesROMS. As interpolated satellite observations and hydrodynamic models are two promising methods for estimating environmental conditions in the absence of in situ networks, a comparison of relative performance in the Chesapeake Bay is instructive both for independent accuracy assessment and for evaluation of data merging potential. Results show that for salinity, ChesROMS tends to perform better, on average, than both OK and UK interpolation of MODIS-derived estimates at most sampling stations throughout the year. Conversely, for temperature, results show that interpolation, on

average, tends to outperform ChesROMS at the majority of the 12 in situ stations throughout the year (Table 2.4). This was not surprising as the salinity estimates, on which the interpolations were performed, were based on extremely variable, bio-optical parameters found in hydrodynamically complex regions of the Bay. That being said, the magnitude of the interpolated salinity errors was fairly small relative to the natural variability of salinity in the Chesapeake Bay.

While ChesROMS salinity estimates exhibited lower error than interpolated satellite estimates in the average, this study reveals systematic patterns in relative performance that can inform data assimilation and other data merging exercises. Performance comparisons, based on a two-year average, show seasonal patterns that generally follow cloud-free satellite coverage in the Bay (Figure 3.1). This indicates that interpolation methods could potentially estimate salinity with accuracy that is comparable to, or perhaps slightly better than, ChesROMS during data heavy months of the year. Satellite products have an advantage in that they are able to capture the variability of extreme conditions due to their diagnostic nature. Hydrodynamic models have the potential to misrepresent actual conditions for a variety of reasons, including the fact that the spatial scale of regional weather events extends beyond the local model grid. An illustrative example occurred in September 2003 when Hurricane Isabel passed through the Chesapeake Bay. Further analysis shows that interpolation techniques outperform ChesROMS for salinity at most of the stations sampled (CB5.5, CB6.1, and CB7.2 were not sampled) following the event (Figure 3.8). Erroneously low modeled salinity after the hurricane is likely due to the nudging of climatological values at the open ocean

boundary. Due to the lack of realistic open-ocean boundary conditions, ChesROMS will fail to capture the influx of salt water brought into the Bay by Isabel.

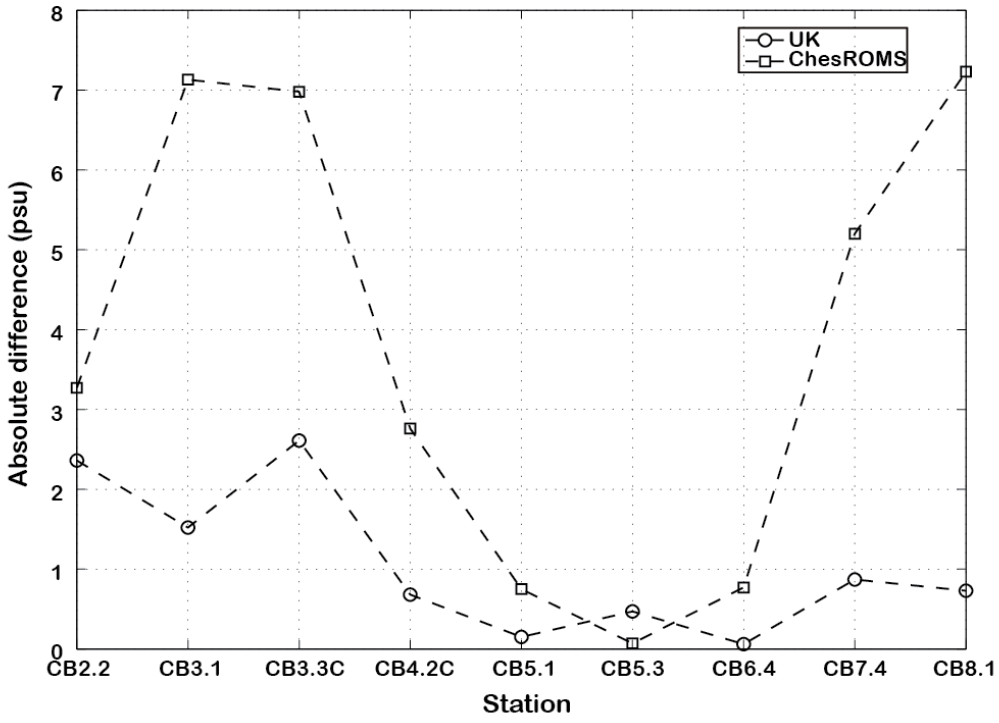

Figure 3.8 MAE between in situ and estimated salinity using UK (empty circles) and ChesROMS (empty squares) following Hurricane Isabel in September 2003

Additionally, there is a potential benefit of using interpolated satellite products to aid in data assimilation efforts to improve performance of models such as ChesROMS in the Bay. Interpolated parameter estimates can be used as observational measurements for the whole Bay for data assimilation. Large errors between interpolated and observed values do not necessarily hinder potential evaluation efforts due the fact that data assimilation accounts for errors and weights them accordingly. Such a data assimilation system using

the Local Ensemble Transform Kalman Filter (Hunt et al., 2007) is being developed for the Chesapeake Bay (Hoffman et al., 2012).

With respect to temperature, interpolated satellite estimates had a smaller average error than ChesROMS. Model comparisons of kriging and ChesROMS revealed seasonal trends that loosely mirror the inverse percentage of satellite coverage in the Bay (Figure 3.1 and Figure 3.7). This suggests that during satellite-limited months, ChesROMS can be expected to provide more reliable estimates of SST than interpolated satellite observations. As with interpolated salinity, interpolated temperature has the potential to aid in future data assimilation and data merging efforts with models including ChesROMS. In this way, the observations could be utilized when they are available and reliable, but model dynamics would drive accurate forecasts in the absence of observations.

Beyond these details of relative performance, comparisons of ChesROMS and interpolated satellite estimates of temperature and salinity are inherently valuable because the two methods represent independent approaches to environmental monitoring. Where ChesROMS is a physically-based model informed by meteorological reanalysis, terrestrial inputs, and ocean boundary conditions, the interpolated MODIS estimates are based on radiation measurements (in this case water-leaving radiance in visible wavelengths for salinity and emitted thermal infrared radiation for temperature). Moreover, while in situ measurements and model output can provide information at sub-surface depths, satellite remote sensing is limited by its inability to penetrate below the

ocean's surface. Spatial interpolations of satellite-derived salinity and temperature thus act as a second source of Bay-wide observations that can serve as an independent check on the hydrodynamic model's representation of circulation or the water body's response to an extreme weather event.

The systematic nature of error fields suggests that satellite estimates, rather than interpolation algorithms, were the primary source of error when interpolated satellite estimates were compared to in situ measurements. We do note, however, that kriging methods require the use of Euclidean distance between locations, and that complications can arise when using Euclidean distance in nonconvex regions such as Chesapeake Bay. This is because straight lines between many pairs of locations go over land, causing relationships between points connected by such lines to be profoundly different than relationships not intercepted by a land boundary (Curriero, 2006; Murphy et al., 2010). An alternative to Euclidean distance is being implemented in the Chesapeake Bay, which will enable kriging analysis in nonconvex regions by defining new spatial metrics based on "water distance" (Murphy et al., 2012).

An additional limitation of this study is the comparison of satellite and model data to point in situ measurements. For method comparison and performance evaluation, we assumed that in situ, satellite, and model measurements were all made at the water "surface". However, due to sampling techniques, optical properties, and various model input, the measurement depths varied within the first meter of the water column. The in situ surface salinity and temperature measurements used for performance evaluation in

this study were taken at a water depth of approximately 0.5m. The depth of the satellite salinity observations was dependent on optical properties and the diffuse attenuation coefficient, which averaged to be a depth of 0.89m for all observations. Remotely sensed temperature is a "skin" measurement and therefore was estimated at the air-sea surface interface. ChesROMS surface measurements were represented by the top "layer", which has potential to cause measurement inconsistencies due to varying depths and varying properties within that first layer. Therefore, though observations were dominated by near-surface samples/estimations, there was a slight discrepancy between the depth of the in situ measurements and the estimated measurements. The fact that ChesROMS simulates the entire water column while satellites only observe the water surface is a strong motivation for data assimilation, which can make use of surface satellite observations to update model fields at all depths.

Overall, the work performed here has been able to make use of the ever-increasing amount of satellite data, computer-modeling output, and in situ parameter data to provide a comparison of environmental parameter estimation methods in the Chesapeake Bay. The interpolation methods and results presented in this study will be used in further work on species-specific empirical habitat models in the Chesapeake Bay. Furthermore, the method of merging remote sensing and geospatial interpolation techniques can be applied to other coastal regions to address the issue of limited data availability, and thus improve upon existing environmental monitoring methods.

Table 2.1 Percent of Data Sets that Best Fit Each Possible Variogram Shape

Parameter	Method	Linear (%)	Exponential (%)	Spherical (%)
Salinity	OK	75	12	13
	UK	46	27	29
Temperature	OK	56	18	26
	UK	28	33	39

Table 2.2 Validation of Interpolation Performance Averaged over All Years

Parameter	Method	ME	MAE	RMSE	R^2
Salinity	OK	0.80	2.14	2.82	0.83
(psu)	UK	0.55	1.88	2.39	0.87
Temperature	OK	0.21	0.60	0.79	0.99
(°C)	UK	0.13	0.60	0.79	0.99

Table 2.3 Percent variance explained by the leading two EOF modes

Parameter	Method	1^{st} Mode (%)	2^{nd}Mode (%)	Total (%)
Salinity	OK	79	13	92
	UK	69	21	90
Temperature	OK	66	24	90
	UK	77	17	94

Table 2.4 Validation of Interpolation and Model Performance Averaged over 2003 and 2007

Parameter	Method	ME	MAE	RMSE	R^2
Salinity	OK	1.08	2.26	3.05	0.78
(psu)	UK	0.61	1.89	2.47	0.86
	ROMS	-0.52	1.46	1.86	0.94
Temperature	OK	0.32	0.67	0.84	0.99
(°C)	UK	0.28	0.63	0.84	0.99
	ROMS	-0.76	1.14	1.39	0.98

4. CHAPTER 4: USE OF ENVIRONMENTAL PARAMETERS TO MODEL PATHOGENIC VIBRIOS IN CHESAPEAKE BAY[10]

ABSTRACT

The abundance and distribution of *Vibrio* spp. is increasing throughout the tributaries of Chesapeake Bay. Annual reports show human infections caused by *Vibrio* spp. have nearly doubled over the past decade in Virginia and Maryland. *Vibrio* spp. are autochthonous to estuarine and coastal waters and follow a seasonal cycle attributed mainly to fluctuations in water temperature and salinity. This study describes development of empirical methods to model the likelihood of occurrence and abundance of *Vibrio* spp. in Chesapeake Bay. To model likelihood of occurrence, a set of binary classification models was developed, employing a suite of geophysical predictor variables and statistical methods. Accuracy of results was ~ 68% at 0.40 prediction for *V. vulnificus* and ~70% at 0.60 prediction for *V. parahaemolyticus*. To model bacterial abundance, regression methods were applied to samples positive for *Vibrio*, showing *Vibrio* abundance can be predicted as a function of sea surface temperature and salinity in Chesapeake Bay, with mean absolute error (MAE) of 3.9 cells/10 ml for *V. vulnificus* and 5.8 cells/10 ml for *V. parahaemolyticus.* A two-step classification/regression hybrid approach was used to generate estimates of abundance in the absence of bacteriological data on presence of *Vibrio* spp. in the Bay. This hybrid approach predicted *Vibrio* abundance with MAE of 2.8 cells/10 ml for *V. vulnificus* and 4.4 cells/10 ml for *V.*

[10] Urquhart E.A., Guikema S.D., Zaitchik B.F., Haley B.J., Taviani, E., Chen, A., Brown, M.E., Huq, A., Colwell, R.R. Use of Environmental Parameters to Model Pathogenic *Vibrios* in Chesapeake Bay. *Journal of Environmental Informatics*, *(Accepted).*

parahaemolyticus. It is concluded that the hybrid method can predict both presence and abundance with accuracy equal to or greater than predictive models requiring bacteriological data for presence of *Vibrio* spp.

4.1. Introduction

The microbiology of the Chesapeake Bay includes many species of the family *Vibrionaceae,* some of which are pathogenic to humans and marine animals (Colwell et al., 1977; Hoge et al., 1989; Wright et al., 1996). Cases of human infection are infrequent, but reports from local health departments and the Centers for Disease Control and Prevention indicate the annual number of reported human *Vibrio* infection cases in the Bay region has nearly doubled in the past decade (Maryland Department of Health and Mental Hygiene, 2013; Virginia Department of Health, 2013). Furthermore, *Vibrio* spp. is frequently detected in oysters and other shellfish harvested for human consumption during the summer months (Constantin de Magny et al., 2009). This seasonality correlates with peak incidence of disease caused by *Vibrio* spp. Soft tissue infections, gastroenteritis, and primary septicemia following consumption of contaminated seafood or exposure to the marine environment are the most common manifestations of *V. vulnificus* disease in humans (Howard and Bennett, 1993; Wright et al., 1996; Strom and Paranjpye, 2000). *V. parahaemolyticus* is an invasive bacterium that typically causes severe diarrhea, but can also cause skin infections if wounds are exposed to seawater or contact with shellfish or crustaceans (Howard and Bennett, 1993; Centers for Disease Control and Prevention, 2013).

Despite the fact that *Vibrio* spp. are known pathogens of global occurrence, the environmental conditions associated with risk of *Vibrio* infection are poorly characterized, with no scientific consensus on the effect of climate change on *Vibrio* populations or risk of *Vibrio* infection. In Chapter 5 we examined *V. vulnificus* model sensitivity to climatic variability and change within the upper Chesapeake Bay by assessing model response to a range of temperature and salinity values. Results showed that the predicted response of *V. vulnificus* probability to high temperatures in the Bay differed systematically between models of differing structure, indicating that the impact of climatic change on the probability of *V. vulnificus* presence in the Chesapeake Bay remains uncertain. Development of regionally customized models for monitoring and predicting risk can empower public health authorities in risk management and controlling vibriosis under evolving climate conditions.

In the Chesapeake Bay, where *Vibrio* spp. are an increasing public health concern, many studies (Kaneko and Colwell, 1973, 1974; Colwell et al., 1977; Kaper et al., 1981; Wright et al., 1996; Parveen et al., 2008) have documented the relationship between *V. vulnificus* and *V. parahaemolyticus* and environmental parameters. In general, abundance of *Vibrio* spp. is greatest when the temperature is greater than 15°C, with salinity between 5 and 25 ppt, and optimal conditions varying by species and region. Temperature and salinity requirements for growth of *Vibrio* spp. have been shown to be related to the seasonal *Vibrio* cycles in coastal and estuarine environments (Kaper et al., 1981; Motes et al., 1998; Lipp et al., 2001; Jacobs et al., 2010).

Other environmental variables can also influence the abundance and distribution at seasonal and subseasonal scales. Yamazaki and Nwadiuto (2012), showed a positive correlation between the concentration of *Vibrio* spp. in coastal waters off the southeast coast of Florida and rainfall, concluding that the decrease in salinity, increased eutrophication, and increased turbidity from terrestrial runoff after rain events were responsible for the observed increase.

Environmental parameters related to the abundance of *Vibrio* spp. and plankton in Chesapeake Bay have been studied extensively (Wright et al., 1996; Louis et al., 2003; Constantin de Magny et al., 2009; Jacobs et al., 2010; Parveen et al., 2013). With the goal of modeling the presence of *V. cholerae* as a function of environmental factors in the Chesapeake Bay, Louis *et al.* (2003) developed an empirical habitat model using logistic regression and a binary classification tree. They showed variations in sea surface temperature and salinity contribute to variability in both frequency of bacterial occurrence and geographic distribution of *V. cholerae*. Wright et al. (1996) and Jacobs et al. (2010) developed similar predictive models for presence of *V. vulnificus,* using *in situ* temperature, salinity, and sampling depth data and logistic regression analysis (Wright et al., 1996) in the Bay. Parveen et al. (2013) developed a predictive model, using temperature, salinity, harvest season, and region on the growth rate of *V. parahaemolyticus* in oysters in the Chesapeake Bay.

Long-term hindcasts and forecasts from predictive models of *Vibrio* spp. can be useful in understanding how land-use and climate change impact the frequency, distribution, and

magnitude of bacteria in the Chesapeake Bay. The information can then be applied to long-term projections of *Vibrio* spp. in the Bay.

Satellite remote sensing, interpolated-satellite, and simulated hydrodynamic model data can be used to achieve temporal and spatial *Vibrio* spp. predictions for the Bay. In fact, a previous study by Constantin de Magny et al. (2009) successfully generated spatially complete predictions of *V. cholerae* likelihood that was based on simulated sea surface temperature and salinity from the numeric model Chesapeake Bay Regional Ocean Modeling System (ChesROMS; (Xu et al., 2012)). Hindcast prediction, distribution and potential hotspot of occurrence of *V. vulnificus* in the Chesapeake Bay has been reported by using a multivariate habitat suitability model stimulated by sea surface temperature and salinity during a period of 1991 and 2005 (Banakar et al., 2011). Banakar *et al.* (2011) concluded that hindcast prediction should be useful for further understanding of the impact of environmental conditions in the occurrence of *V. vulnificus* and long-term projections of *Vibrio* spp. in the Chesapeake Bay. Thus, satellite and *in situ* observations can be combined in a dynamical model with data assimilation so that observations when available are utilized, and the model dynamics drive forecasts in the absence of observations. Furthermore, a data assimilation system, using ChesROMS, has recently been developed (Hoffman et al., 2012) for the Chesapeake Bay.

Here we present empirical algorithms for predicting the probability of *Vibrio* spp. incidence and abundance in the upper Chesapeake Bay, which represent an advance over existing models in two respects. First, a model for *V. parahaemolyticus* presence and

concentration in Chesapeake Bay is provided. Second, concentration of *Vibrio* spp. in areas where they are present can be obtained. Since the risk of human infection is a function of *Vibrio* concentration, extending available predictive models to provide concentration, in addition to presence/absence, advances the public health utility of the models significantly.

In effect, this study contributes to environmentally-based pathogen prediction by incorporating a range of statistical modeling options. Most ecological forecasting models rely on a single model structure, usually linear regression. In contrast the current study tests three types of empirical model: Generalized Linear Model (GLM), Generalized Additive Model (GAM), and Random Forest Model (RF). In using the three models, we have taken a multistep approach: first, binary classification is used to model whether or not bacteria are present; second, regression of positive count data is used to estimate bacterial abundance; third, the methods are combined using hybrid classification-regression, estimating total bacterial abundance in a given geographic area predicted to have *Vibrio* spp. present. Thus, the main objectives of this study were to develop a *Vibrio* spp. empirical algorithm capable of producing likelihood of presence maps and develop a *Vibrio* spp. algorithm that estimates bacterial abundance in a given geographical area of the Chesapeake Bay.

4.2. Materials and Methods

4.2.1. Sample collection

Water samples were collected during July and September, 2011, and March through June, 2012, at sites located in Chesapeake Bay (See Figure 4.1). The Maryland Department of Natural Resources and the NASA GEO-CAPE Field Campaign research vessels were used in the sampling, with surface water samples (0.5-1 m depth) collected using a combination of flow-through collection systems and overboard bucket sampling. For the latter, sterile polypropylene bottles (1 L) were rinsed, filled, and placed on ice for transport to the laboratory within 1 hour of collection. Surface temperature and salinity were measured at the time of collection of water samples using an YSI Series 6 instrument (Yellow Springs, Ohio). A total of 148 surface water samples were collected for bacteriological analysis that was carried out within 12 hours at the Maryland Pathogen Institute located at the University of Maryland, College Park.

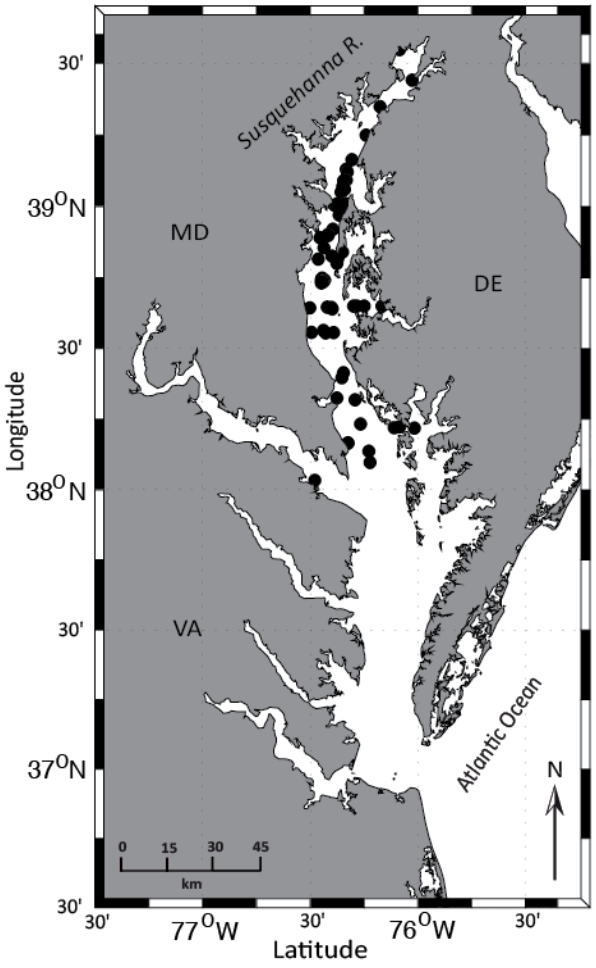

Figure 4.1 Map of Chesapeake Bay and its tributaries: dark circles represent the sampling stations for this study.

4.2.2. Laboratory Sample processing

4.2.2.1. DNA Extraction and Qualitative Direct PCR

Water samples were shaken and 100 ml passed through a 0.22 μm sterile polycarbonate membrane, then placed in 5 ml of sterile 1X PBS and vortexed. A 1 ml aliquot was removed and boiled for 10 minutes, and iced for 10 minutes before centrifuging at 13,000 rpm for 10 minutes. The supernatant was transferred to a sterile microcentrifuge tube and

stored at 20^{o}C until *toxR* multiplex PCR was employed for qualitative detection of *V. vulnificus*, and *V. parahaemolyticus* (Bauer and Rorvik, 2007). Results were visualized on 1% agarose gel stained with ethidium bromide.

4.2.2.2. Quantitative Colony Blot Hybridization

To quantify culturable *V. parahaemolyticus* and *V. vulnificus* plates, 1 ml water samples were spread in duplicate onto T1N3 agar and *Vibrio vulnificus* agar (VVA) plates, respectively, and the plates were incubated overnight at 37°C. Colonies were lifted onto Whatman #541 filters and species-specific probe hybridization was done (DePaola et al., 1997; McCarthy et al., 1999).

4.2.3. Statistical Model

Three statistical modeling methods were used, Generalized Linear Modeling (GLM), Generalized Additive Modeling (GAM) and Random Forest models (RF) to predict three characteristics of *Vibrio* spp. distribution, namely probability of presence (hereafter: "LIKELIHOOD"), abundance at sites with confirmed presence (hereafter: "ABUNDANCE"), and abundance at all sites in the absence of prior bacteriological data on presence (hereafter: "HYBRID," because it requires a two-step classification/regression approach). ABUNDANCE models assume perfect prior information on presence/absence and were included to determine how models would perform in addressing presence versus absence and quantitative prediction of bacterial abundance. The HYBRID models provide prediction and offer realistic operational potential.

All statistical computations were carried out using R Statistical Package 2.14 on an Intel Xeon W3580 Processor, 3.33 GHz machine with 12 GB RAM. Computation time for all likelihood statistical models within the holdout validation test was less than three minutes.

4.2.3.1. Statistical methods

The GLM, GAM, and RF modeling methods were used to develop LIKELIHOOD, ABUNDANCE, and HYBRID models. For LIKELIHOOD models, each method was implemented in logistic form and trained using observational data transformed to binary presence/absence: cell count > 0 cells/10 ml ≡ presence, cell count = 0 cells/10 ml ≡ absence. For ABUNDANCE models, cell count was predicted as a continuous variable. The ABUNDANCE models were developed using data only from samples with cell counts > 0, and a log link function was applied in GAM and GLM, using a Poisson likelihood function. HYBRID modeling was carried out using a two-step technique described by Guikema and Quiring (2012): (1) binary classification based on the best LIKELIHOOD model, (2) concentration prediction based on the best ABUNDANCE model.

4.2.3.1.1. Generalized Linear Model (GLM)

The Generalized Linear Model is an extension of the Ordinary Least Squares (OLS) linear model that allows for non-Gaussian probability distributions and the use of both continuous and count data (Nelder and Wedderburn, 1972; Fox, 2008). GLM achieves

flexibility by including a link function that relates linear predictor to a function of the explanatory variables (Cameron and Trivedi, 2013). For binary data, one such function is the "logit" link function and it transforms expectation of response to the linear predictor:

$$log\,[p / (1 - p)] = \beta_0 + \Sigma_j \beta_j x_j\,, \quad (1.3)$$

where $p / (1 - p)$ is the odds ratio of *Vibrio* spp. presence, β_0 is the intercept, β_j is the regression coefficient for variable x. Furthermore, solving for p, the probability of *Vibrio* presence is then:

$$P_{\text{presence}} = e^{(logit)} / [e^{(logit)} + 1]\,. \quad (1.4)$$

The GLM algorithm was implemented by the *stats* (version 2.14.0) R package (Hastie and Pregibon, 1992).

4.2.3.1.2. Generalized Additive Model (GAM)

A Generalized Additive Model extends GLM by allowing for nonlinear relationships between explanatory variables and response variable (Hastie and Tibshirani, 1990). This is achieved by replacing the linear predictor $\alpha + \Sigma_j \beta_j x_j$ of a GLM with an additive predictor $\alpha + \Sigma_j f_j(x_j)$ where $f_j(x_j)$ is a non-parametric smoothing function. The smoothing function provides information about the relationship between explanatory variables and response variable not revealed using a traditional linear model (Hastie and Tibshirani, 1986). For this study, the standard smoothing approach, a cubic regression spline, was used. Again, for bacterial presence data, the "logit" link function was used to establish the relationship between response variable and smoothed function of the explanatory variables. The GAM algorithm was implemented by the *mgcv* (version 1.7-16) R package (Wood, 2006).

4.2.3.1.3. Random Forest (RF) Model

A Random Forest model is an algorithm that fits many classification trees to a dataset, and then uses an ensemble of tree-structure predictions (Breiman, 2001). The algorithm begins with selection of *n* bootstrapped samples (e.g., 500) with replacement from the original dataset. Observations from the original dataset not included in the bootstrap sample are referred to as out-of-bag (OOB) sample, and are used in model cross-validation. A classification tree is fit to each bootstrap sample. To ensure that each of the trees in the ensemble is independent, each tree uses a small number (*m*) of randomly selected predictor variables for split construction at each node. The trees are fully grown and each individual tree is used to estimate the OOB sample. The predicted class is calculated by a majority vote of the OOB predictions for that sample. The RF algorithm in this study was implemented by the *randomForest* (version 4.6.-6) R package (Liaw and Wiener, 2002).

4.2.3.2. Model evaluation

4.2.3.2.1. LIKELIHOOD model validation

Predictions from the LIKELIHOOD models come in the form of probabilities, such that a probability threshold or prediction point is needed to transform probability into bacterial presence/absence data. A prediction point is also required to assess model performance using various indices derived from a confusion matrix. Rather than subjectively setting probability to an arbitrary value of 0.50 (50%), which has no ecological basis, the threshold was selected empirically to maximize agreement between observed and

predicted distributions in the out of bag data. To ensure correct binary classification, we optimized this prediction point relative to four model assessment indices: true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), and accuracy (ACC). In addition, area under the curve (AUC) was calculated for each threshold probability. The indices listed above require information from the confusion matrix, which consists of four elements: true positive (TP), false positive (FP), false negative (FN), and true negative (TN). The indices used to assess the predictive performance of the various LIKELIHOOD models are described below:

$$TPR = TP / (TP + FN), \quad (1.5)$$

where true positives represent bacterial presence predictions and false negatives represent bacteria present but predicted by the model as absent.

$$TNR = TN / (FP + TN), \quad (1.6)$$

where true negative is correctly predicted bacteria presence, and false positives are bacteria absences classified as present by the model. TPR and TNR are widely referred to as sensitivity and specificity; both are used in the Receiver Operator Characteristic (ROC) curve (i.e. sensitivity vs. 1-specificity) whose tangent slope is equal to 1 (Hanley and McNeil, 1982).

$$FPR = FP / (FP + TN), \quad (1.7)$$

where FPR is equivalent to "fall out" which in binary classification is equal to (1-specificity).

$$ACC = (TP + TN) / (P + N), \quad (1.8)$$

where P is the number of actual presence instances and N is the number of absence instances. Selection of the final prediction points was based on a combination of the indices and is explained in detail below.

4.2.3.2.2. ABUNDANCE and HYBRID model evaluation

The predictive accuracy of *Vibrio* spp. ABUNDANCE models was assessed using random holdout validation analysis. Datasets for each species were randomly partitioned into a training dataset containing 80% of the original records and a validation dataset containing the remaining 20%. The models described above were developed using the training dataset and subsequently employed to predict cell number using the holdout dataset. This process was repeated 100 times with a different random partition each time. Mean error (ME) and mean absolute error (MAE) were used to compare estimated bacterial abundance to observed abundance, identify outliers in each model fit, and evaluate comparative model performance.

HYBRID models were evaluated with the same presence-only validation dataset used to assess the ABUNDANCE models. The hybrid models were assessed with presence-only holdout dataset, denoted "HYBRID/P". To measure hybrid method performance in predicting *Vibrio* spp. abundance at all sample locations, without bacteriological data input, the original validation dataset containing both zero and non zero records was used. "HYBRID" denotes hybrid models evaluated with original holdout dataset. Additionally, unweighted model averages were calculated for both species. All hybrid analyses employed the LIKELIHOOD model structures shown in Table 3.3.

4.2.3.2.3. Mean model

ABUNDANCE and HYBRID models were compared to a mean statistical null model, i.e., the average value of the response variable, *Vibrio* spp. For validation, empirical models including the mean model were input to the holdout analysis.

4.3. Results

4.3.1. Observations

Over the eight months during which *Vibrio* spp. counts in the water samples were obtained, 46% contained *V. vulnificus* and 68% contained *V. parahaemolyticus*. In samples positive for *V. vulnificus*, the median and mean counts were 4 and 6 cells/10 ml respectively, and concentrations ranged from 1 to 30 cells/10 ml (Figure 4.2). For *V. parahaemolyticus*, the median and mean count was 7 and 9.5 cells/10 ml, respectively, and concentrations ranged from 1 to 50 cells/10 ml (Figure 4.2). Counts were obtained for samples collected at temperatures ranging from 8 to 31°C and 0 ppt to 14 ppt salinity. The highest number of *Vibrio* spp. were in water samples at 28°C and salinity of 11.5 ppt (Figure 4.3). These results are consistent with those reported for *Vibrio* spp. in Chesapeake Bay by Jacobs et al. (2010).

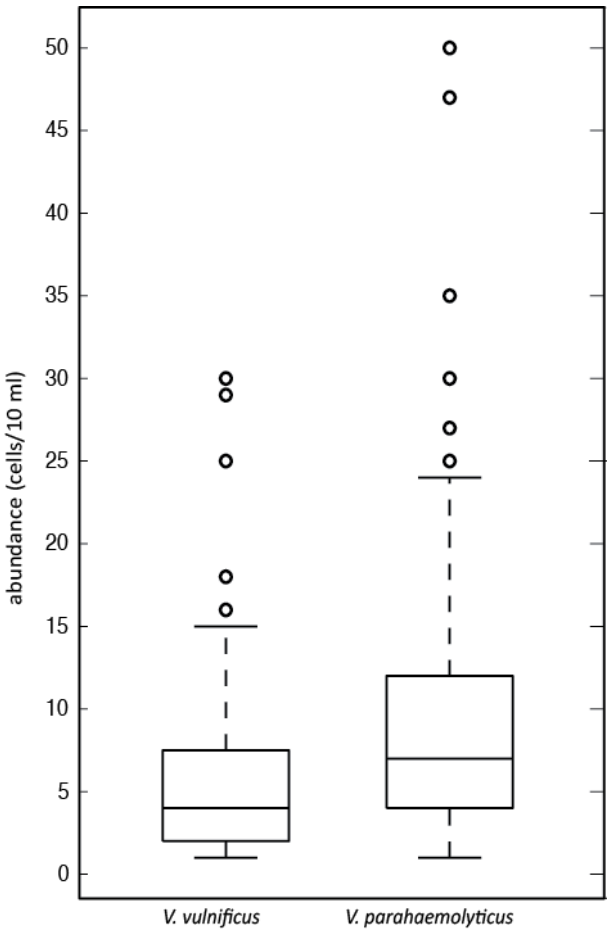

Figure 4.2 Boxplot showing concentration (cells/10 ml >0) for *V. vulnificus* (n=68) and *V. parahaemolyticus* (n=100). Horizontal lines are median cell counts (*V. vulnificus*=4 cells/10 ml and *V. parahaemolyticus*=7 cells/10 ml); boxes represent 25th and 75th percentile and whiskers 5th and 95th percentile. Individual open circles beyond the whiskers represent outliers.

4.3.2. Modeling occurrence and abundance of *Vibrio* spp. in Chesapeake Bay

Descriptive correlation analyses relating environmental predictors to *Vibrio* spp. distribution and results of LIKELIHOOD, ABUNDANCE, and HYBRID predictive models are presented as follows.

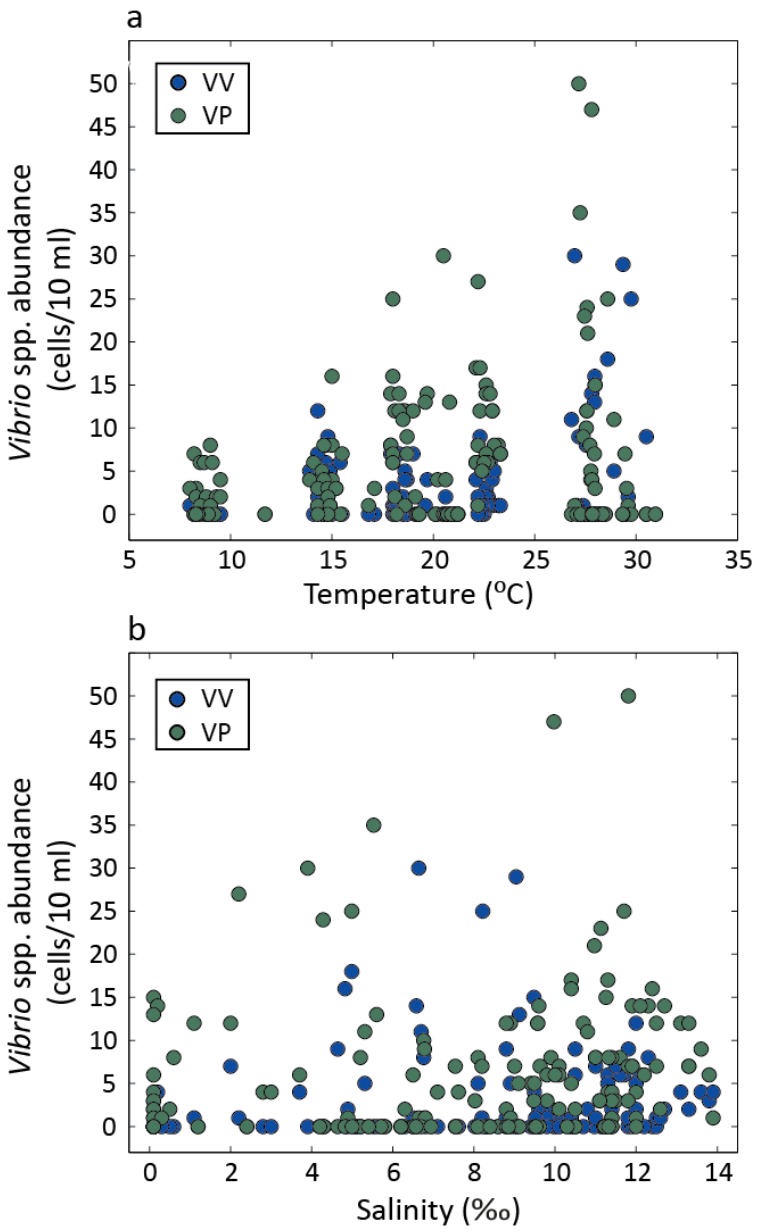

Figure 4.3 Plots shwowing the relationship between counts of *Vibrio* spp. (per 10 ml) and a) temperature and b) salinity.

4.3.2.1. Correlation of *Vibrio* spp. with Environmental Predictors

The predictive potential of environmental parameters, was examined using univariate correlation analysis for *Vibrio* counts in samples containing given the *Vibrio sp*.

Statistically significant correlations were found between *Vibrio* count and surface water temperature, month, and salinity x temperature interaction (Table 3.1). Although statistically significant, the correlation coefficients were low. It is important to note that correlations observed for month and, potentially, interaction may derive from cross-correlation with surface water temperature. For most sampling locations, total bacterial count followed a seasonal pattern following the temperature. Linear correlations between *Vibrio* count and salinity, latitude, or longitude were not statistically significant (Table 3.1). Since freshwater discharge impacts both nutrient inflow and sediment transport, it can influence bacterial abundance.

4.3.2.2. LIKELIHOOD models

A stepwise selection process was used to select a LIKELIHOOD model, whereby each explanatory variable was entered sequentially into each model. The entire suite of models was tested, and selected variables retained only if significant. For the model evaluation, significance was set at an alpha level of 0.05. GLM and GAM logistic regression for both *V. vulnificus* and *V. parahaemolyticus* showed temperature and salinity, and for *V. vulnificus* interaction between the two variables, were core explanatory parameters for the three LIKELIHOOD models. Table 3.2 presents the best-fit models developed for *V. vulnificus* and *V. parahaemolyticus*, where probability of bacteria presence ($P_{\text{presence)}}$ is defined in Equation 1.4.

Figure 4.4 illustrates the probability of *V. vulnificus* being present as predicted by best-fit a) GLM, b) GAM and c) RF LIKELIHOOD models (Table 3.2). Likelihood of presence

was split into absence (n= 48; median prob. = 0.47, 0.26 and 0.27) and presence (n=50; median prob. = 0.56, 0.67 and 0.57) observations. Figure 4.5 shows the probability of *V. parahaemolyticus* predicted by best-fit a) GLM, b) GAM and c) RF LIKELIHOOD models (Table 3.2). Likelihood of occurrence was split into absence (n= 82; median prob. = 0.52, 0.55 and 0.48) and presence (n=40; median prob. = 0.69, 0.78 and 0.87) observations. Points falling outside of the 95th percentile in the boxplots represent outliers.

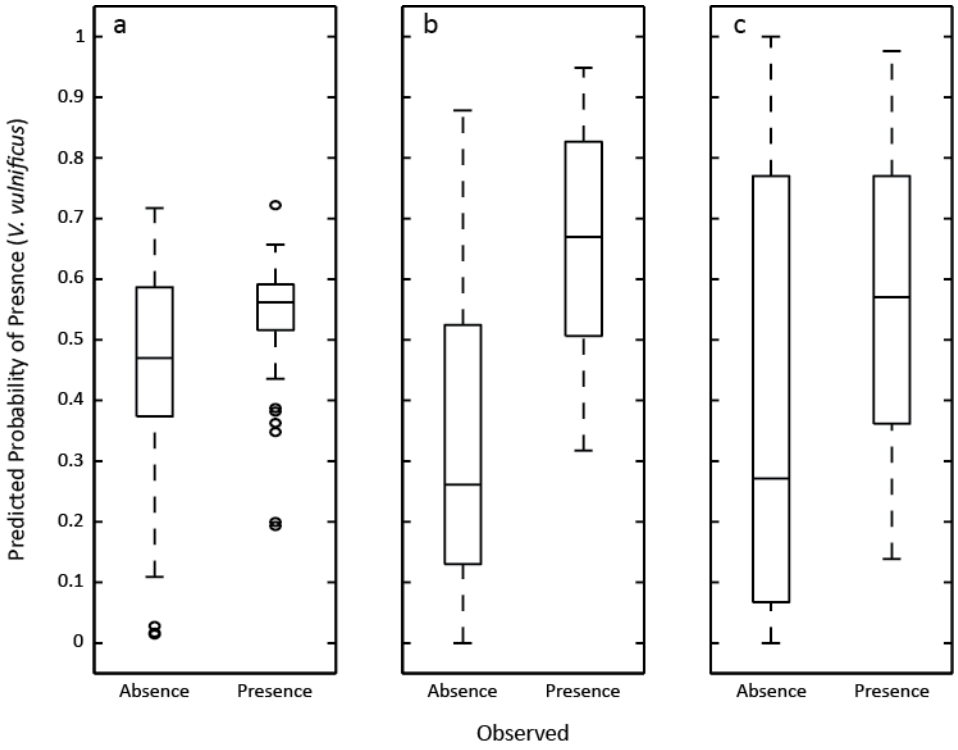

Figure 4.4 Performance of a) GLM, b) GAM and c) RF *Vibrio vulnificus* classification models (Table 3.2), presented as boxplots comparing presence and absence with modeled probability, where threshold for presence is 1 cell/10 ml. Horizontal lines represent median probabilities, boxes the 25th and 75th percentiles, and the whiskers are 5th and 95th percentiles. Open circles beyond whiskers represent probabilities outside the IQR.

LIKELIHOOD GLM, GAM and RF models used to predict bacterial presence required selection of an optimal prediction point or threshold. Rather than setting a prediction point 0.5 arbitrarily, the prediction point for each species was based on four performance indices: TPR, FPR, TNR and ACC. With the goal of maximized model prediction skill and binary classification, information from each of these metrics (Figure 4.6), as well mean and median statistics from predicted probabilities (Figure 4.4), was used to select the optimal prediction point for each species. Because no significant difference was observed between the accuracy index for 0.4, 0.5, and 0.6 prediction points for *V. parahaemolyticus*, each threshold was tested in the holdout analysis, yielding greater accuracy, with an optimal threshold of 0.6. With this information, maximum ACC and TPR were selected, yielding an optimal threshold of 0.4 for *V. vulnificus* (ACC: 0.63 for GLM, 0.72 for GAM, and 0.68 for RF; Table 3.3), and 0.6 for *V. parahaemolyticus* (ACC: 0.62 for GLM, 0.65 for GAM, and 0.67 for RF; Table 3.3).

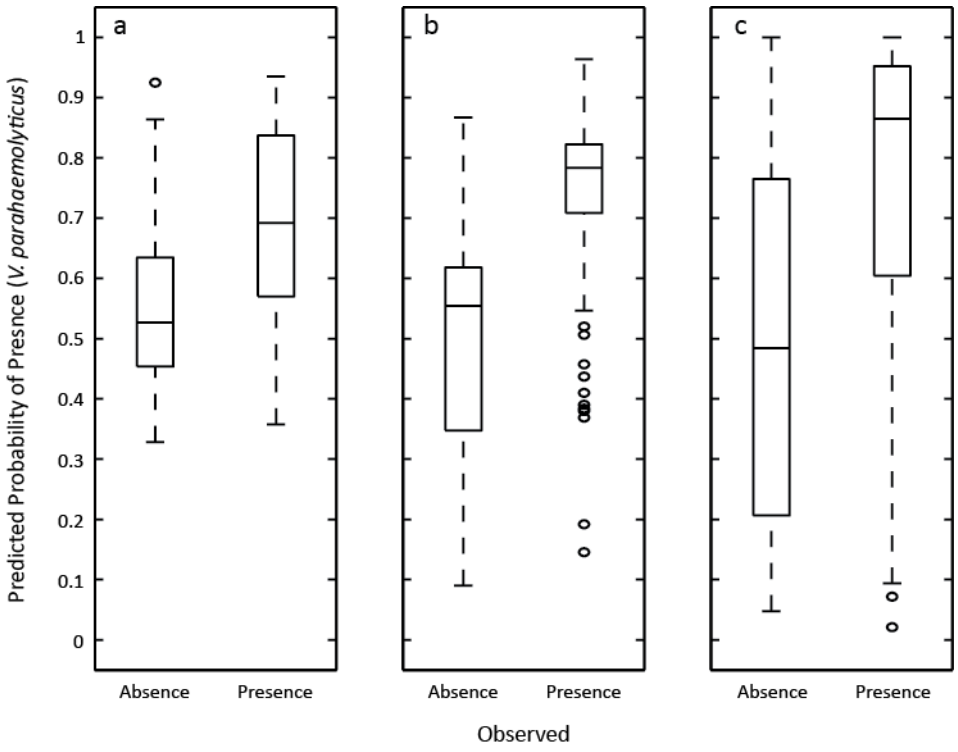

Figure 4.5 Performance of a) GLM, b) GAM and c) RF for *Vibrio parahaemolyticus* classification models (Table 3.3), presented as boxplots comparing presence and absence with modeled probabilities where the threshold for presence is 1 cell/10 ml. Horizontal lines represent median probabilities, boxes the 25th and 75th percentile and whiskers the 5th and 95th percentile. Open circles beyond the whiskers represent probabilities outside the IQR.

4.3.2.3. ABUNDANCE models

ABUNDANCE models described in section 4.2.3.1 were applied to all samples with *Vibrio* greater than 0 cells/10 ml, using repeated random holdout validation tests. Results indicate RF offered better prediction when the bacterial counts were high and GAM and GLM offer better prediction when the numbers were low. Based on these performance patterns, unweighted model average predictions of GAM and RF were tested. For each species, four ABUNDANCE models were then applied: (1) GLM, (2) GAM, (3) RF, (4) model average. Each model was also compared to the mean prediction model in the holdout test to determine how well each model performed relative to assuming the mean *Vibrio* bacterial count for each species, which provides an estimate of the degree to which each empirical model offers an improvement over using the historic mean as the future prediction. This resulted in 10-pair wise tests. Applying the Bonferroni correction for multiple hypothesis tests, a p-value below 0.005 (p=0.05 overall) indicates statistical significance for any given test.

As shown in Table 3.4, the RF ABUNDANCE model provides the best predictive accuracy for *V. vulnificus,* with lowest MAE (3.87 cells/10 ml) followed by average ABUNDANCE MAE (3.94 cells/10 ml). The MAE values were statistically significantly lower than GLM and GAM MAE ($p < 0.005$). The model average and RF model had lower error than the mean model by a statistically significant amount ($p = 0.005$). For *V. parahaemolyticus*, the model average (5.62 cells/10 ml) and RF (5.76 cells/10 ml) had the lowest MAE values, and were lower than MAEs of the GLM and GAM by a statistically significant amount ($p < 0.005$). The difference between MAE for model average and RF were not statistically significant ($p = 0.38$). While all four models were statistically different from the mean predictions, only the model average and RF model outperform the mean model ($p < 0.005$).

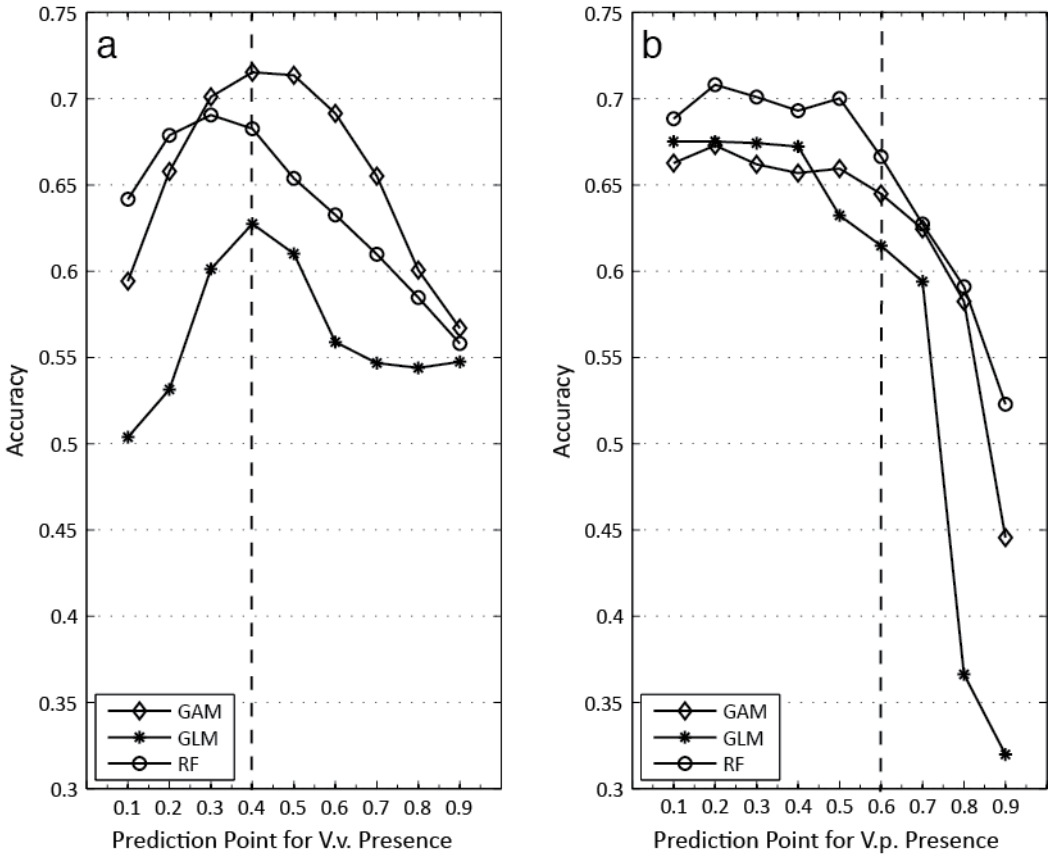

Figure 4.6 Optimization of prediction point (expressed as decimal fraction) to determine p for a) *V. vulnificu*s and b) *V. parahaemolyticus*. Vertical lines indicate optimized prediction point at 0.40 and 0.60 for *V. vulnificus* and *V. parahaemolyticus*, respectively.

The prediction accuracy of each model was examined whereby the predictions were binned based on the actual cell number obtained from the validation datasets (cells/10 ml = 1, 2-4, 5-10 and >10 for *V. vulnificus*) (Figure 4.7a and 4.7c) and (cells/10 ml= 1, 2-4, 5-10, 11-15 and >15 for *V. parahaemolyticus*) (Figure 4.7b and 4.7d). While the RF model had a lower overall MAE than GLM and GAM for *V. vulnificus*, it exhibited a larger MAE in the lower concentration bins (Figure 4.7c) due to over prediction in those bins (high ME values) (Figure 4.7a). For *V. parahaemolyticus*, overall ME values showed all models, except RF, under predicted the cell count because of significant under prediction at high concentrations (Figure 4.7b). The GLM and GAM exhibited lower MAE values at lower concentrations than the RF (Figure 4.7d). However, at counts higher than 5 cells/10 ml, the RF model outperformed both GLM and GAM. Averaging model predictions reduced overall RF MAE, but increased the MAE when counts were high.

4.3.2.4. HYBRID models

Based on error results of both LIKELIHOOD and ABUNDANCE models, a two-step GAM classification/RF regression HYBRID modeling approach was used. Other classification/regression model combinations (e.g., GLM/GLM, GAM/GAM, and RF/RF)

were also tested, but GAM/RF was the best performing hybrid combination. The GAM/RF combination exhibited significantly lower error ($p<0.005$) than other HYBRID combinations for *V. parahaemolyticus*, and similar error for *V. vulnificus*, although the difference was not statistically significant ($p = 0.005$). To assess the prediction accuracy of our HYBRID approach we evaluated the model using two different holdout datasets: (1) a presence-only validation dataset, and (2) the original validation dataset irrelevant of presence or absence. Using the same presence-only validation dataset that was used to evaluate the ABUNDANCE models allowed direct comparison of the prediction accuracy of the HYBRID and ABUNDANCE models. Model evaluation using the original holdout dataset allowed an estimation of *Vibrio* counts without the bacteriological data.

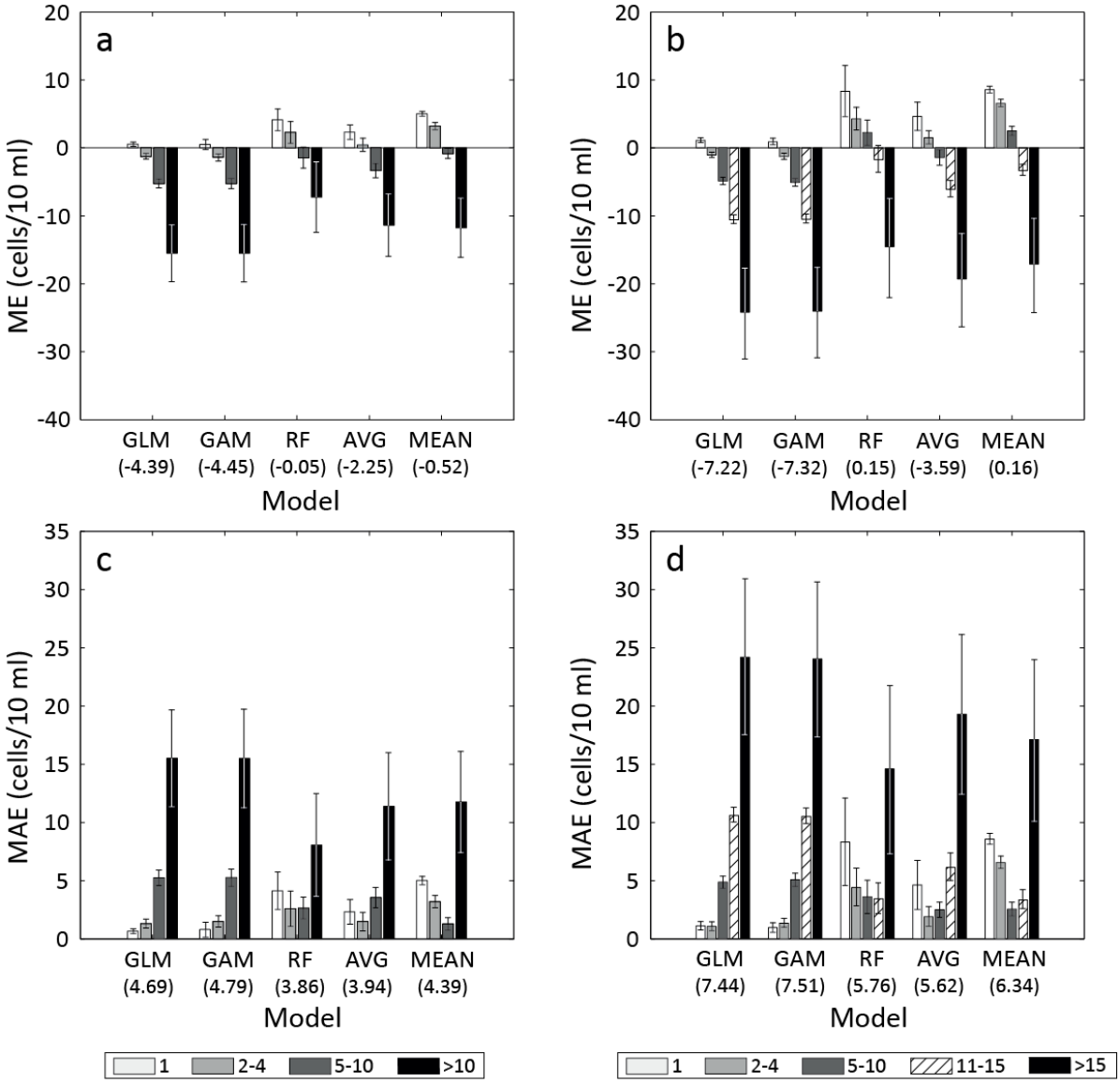

Figure 4.7 Binned ME and MAE values (cells/10 ml) of each ABUNDANCE model for *V. vulnificus* (a & c) and *V. parahaemolyticus* (b & d), shown as bar graphs with error bars (standard deviation). Numbers in brackets represent total error (cells/10 ml).

Table 3.5 compares ME and MAE from the ABUNDANCE RF model with those of the HYBRID model, using the presence-only validation dataset (HYBRID/P), and the model validated with the original dataset (HYBRID) for *V. vulnificus* and *V. parahaemolyticus*.

The RF ABUNDANCE, HYBRID/P, and HYBRID all exhibited negative bias (ME) due to under prediction of counts at high concentrations. Results of the hypothesis tests using the MAE as the error measure showed that when using presence-only data, significant error reduction from the ABUNDANCE RF model MAE (3.9 cells/10 ml) to the HYBRID/P MAE (2.8 cells/10 ml) was observed. When the HYBRID approach was used to predict the original zero and non-zero dataset, an improvement in error relative to ABUNDANCE model was also noted. These two predictions are not exactly comparable, as the ABUNDANCE model was trained and evaluated using only samples with confirmed *Vibrio* counts, while the HYBRID prediction applied to all data, without bacteriological laboratory data.

For *V. parahaemolyticus*, both HYBRID/P and HYBRID exhibited negative bias largely due to under estimation when counts were high, and a positive bias for RF ABUNDANCE model. Both HYBRID/P (4.4 cells/10 ml) and HYBRID (5.26 cells/10 ml) offer an improvement in MAE relative to ABUNDACNE (5.8 cells/10 ml) model. All of the HYBRID models offer improvement over using the mean of the validation dataset.

4.4. Discussion and Conclusions

The empirical models presented in this study demonstrate significant skill in estimating probability of occurrence of *Vibrio* spp., as well as bacterial counts in Chesapeake Bay water samples when bacteriological count data are included. When the HYBRID approach was used to generate estimates of *Vibrio* counts, an overall reduction in error

was observed compared to presence-only ABUNDANCE models when the models were evaluated only at sites with detectable *Vibrio* counts. The fact that HYBRID outperformed ABUNDANCE, when evaluated at sites where *Vibrio* was present, is surprising, since the ABUNDANCE models benefited from data on presence versus absence. Examination of partial dependence plots indicated differences in model performance are a product of differences between the HYBRID and ABUNDANCE models at moderate values of temperature and salinity. The differences were relatively small, however, and no major difference was noted in structure between RF ABUNDANCE model and RF component of the HYBRID model. We conclude that the enhanced performance of HYBRID, relative to ABUNDANCE, most probably derives from the models having been trained for a broader range of conditions—as was the case for HYBRID, since errors in the GAM PRESENCE model led to a more diverse training set for the RF component of the HYBRID—and tend to be more generalizable than models trained under more narrow conditions, even when these narrow conditions capture the range of the specific response variable of interest (Nateghi and Guikema, 2013). HYBRID performed at least as well as ABUNDANCE, which indicates that the HYBRID approach allows for modeling both presence and abundance without loss of skill relative to an abundance model supplied with perfect prior information on presence versus absence.

When both the HYBRID and ABUNDANCE models were evaluated for all sites, the predictive accuracy of the HYBRID was better than that of the ABUNDANCE model, though the difference was not statistically significant for *V. parahaemolyticus.* Similar to

model behavior observed for the ABUNDANCE models for both species, overall error reduction using the HYBRID modeling approach showed the two-step approach tends to over predict counts at low *Vibrio* concentrations. Furthermore, when evaluating prediction performance of each model relative to the mean model, a statistically significant improvement over the mean value of the validation dataset in all models was noted, except for the *V. parahaemolyticus* HYBRID model. It is important to emphasize that when using the complete original dataset for validation (HYBRID), zero-inflation and a lower overall mean model value must be considered. In future model evaluation using zero-inflated datasets, alternative methods of mean model comparisons should be employed.

The empirical models presented in this study offer a novel approach for estimating *Vibrio* spp. abundance in Chesapeake Bay water. We note that the study was limited by the small number of samples available to train and evaluate the models. First, the field data used in this study was limited to the oligohaline (0-6 ppt) and mesohaline (6-18 ppt) regions of the upper Chesapeake Bay. Because our models were trained using data for fresh and brackish water, extrapolation, of the models to saline regions may result in greater error and thus, decreased accuracy of prediction for *Vibrio* spp. near the mouth of the Chesapeake Bay. Specific attention to this discrepancy will be required if the models developed in this study are to be applied to coastal regions. Therefore, data from more saline waters will be needed to train the model. Work in progress covers whole Bay hindcast predictions using temperature and salinity from satellite sensors to understand long-term trends of *Vibrio* spp. likelihood and abundance throughout the Bay. Water

samples were collected only in the upper Chesapeake at a limited number of stations over a two-year period. A longer and more intense sampling record would be valuable to produce more robust models with improved predictive capability.

In summary, we have presented several empirical algorithms for estimating the likelihood of *Vibrio* occurrence as well as abundance (cells/10 ml) in Chesapeake Bay surface water. To estimate the probability of *Vibrio* spp. being detected in Bay water, we tested several binary classification methods. To model *Vibrio* spp. abundance, several regression methods were applied to samples positive for *Vibrio* spp. A two-step hybrid approach using GAM for classification and RF for regression was employed to estimate abundance of *Vibrio* spp. in the absence of bacteriological data. For LIKELIHOOD models, GAM demonstrated a greater accuracy and improved positive rate than GLM and RF models. ABUNDANCE models, GLM and GAM exhibited higher prediction accuracy when counts of *Vibrio* spp. were low. However, RF exhibited lower overall mean absolute error. HYBRID performed better than ABUNDANCE at sites where *Vibrio* presence had been confirmed by bacteriological methods, and predicted abundance as well or better than ABUNDANCE even when evaluated for sites both with and without *Vibrio* spp. confirmed to be present. Thus, HYBRID modeling offers the potential to predict both presence and abundance of *Vibrio* bacteria in Chesapeake Bay surface water. Since presence and abundance of *Vibrio* spp. are relevant to the risk of infection, this capability offers meaningful improvement over existing monitoring and prediction systems.

Table 3.1 Correlation coefficients for *Vibrio* spp. counts and list of selected environmental variables. Significant correlation at the alpha=0.05 level is highlighted in **bold**.

	VP	Lat	Lon	Month	Temp	Saln	Inter
VV (cells/10 ml)*	0.04	-0.03	-0.11	**0.25**	**0.28**	0.09	**0.22**
VP (cells/10 ml)*		-0.10	0.01	0.13	**0.17**	0.09	**0.19**
Latitude			0.15	0.03	0.06	**-0.75**	**-0.56**
Longitude				0.06	0.06	**-0.32**	-0.15
Month					**0.96**	-0.04	**0.54**
Temperature (^{o}C)*						-0.04	**0.57**
Salinity (ppt)*							**0.76**
Interacton*							

*Included in final model development

Table 3.2 Best-fit LIKELIHOOD algorithms for V. *vulnificus* and *V. parahaemolyticus*, probability of presence ($P_{presence}$) is a function of logit

	V. vulnificus	*V. parahaemolyticus*
Model	$P_{\text{presence}}=e(^{logit})/[e(^{logit}) + 1]$**	$P_{\text{presence}}=e(^{logit})/[e(^{logit}) + 1]$**
GLM	$logit= \beta_0 + \beta_1[\text{T}] + \beta_2[\text{S}]+ \beta_3[(\text{T*S})]$	$logit= \beta_0 + \beta_1[\text{T}] + \beta_2[\text{S}]$
GAM	$logit= \beta_0 + S_1[\text{T}] + S_2[\text{S}] + S_3[[\text{T*S})]$	$logit= \beta_0 + S_1[\text{T}] + S_2[\text{S}]$
RF	$P_{presence} = randomForest(\text{T} + \text{S} + (\text{T*S}))$	$P_{presence}= randomForest(\text{T} + \text{S})$

**not applicable to RF model

Table 3.3 *V. vulnificus* and *V. parahaemolyticus* (LIKELIHOOD) performance metrics at prediction point 0.40 for *V.vulnificus* and 0.60 for *V.parahaemolyticus*

	V.vulnificus			*V.parahaemolyticus*		
	GLM	GAM	RF	GLM	GAM	RF
AUC	0.68	0.78	0.73	0.63	0.70	0.71
FPR	0.44	0.35	0.37	0.30	0.24	0.22
TPR	0.81	0.81	0.76	0.42	0.48	0.50
TNR	0.56	0.65	0.63	0.70	0.76	0.78
ACC	0.63	0.72	0.68	0.62	0.65	0.67

Table 3.4 Comparison of holdout ABUNDANCE MAEs (cells/10 ml) based on 100 random holdout samples for *V. vulnificus* and *V. parahaemolyticus*. p-Values in bold represent statistically significant differences between models at the alpha=0.005 level.

Model	Mean MAE	GAM	RF	AVG	MEAN
		V. vulnificus			
GLM	4.69	0.61	**<0.01**	**<0.01**	0.09
GAM	4.79		**<0.01**	**<0.01**	0.02
RF	3.87			0.61	**<0.01**
AVG	3.94				**<0.01**
MEAN	4.39				

V. parahaemolyticus					
GLM	7.43	0.70	**<0.01**	**<0.01**	**<0.01**
GAM	7.51		**<0.01**	**<0.01**	**<0.01**
RF	5.76			0.38	**<0.01**
AVG	5.62				**<0.01**
MEAN	6.34				

Table 3.5 Comparison of MEs and MAEs for ABUNDANCE and HYBRID models for *V. vulnificus* and *V. parahaemolyticus*

V. vulnificus			*V. parahaemolyticus*		
Error Metric	Error	MEAN	Error Metric	Error	MEAN
ME.ABUNDANCE	-0.05	-0.05	ME.ABUNDANCE	0.14	0.16
ME.HYBRID/P	-1.58	-3.25	ME.HYBRID/P	-1.93	-2.98
ME.HYBRID	-0.28	0.19	ME.HYBRID	-1.91	-0.11
MAE.ABUNDANCE	3.87	4.39	MAE.ABUNDANCE	5.76	6.34
MAE.HYBRID/P	2.79	4.30	MAE.HYBRID/P	4.36	5.83
MAE.HYBRID	2.94	3.44	MAE.HYBRID	5.26	6.12

5. CHAPTER 5: UNCERTAINITY IN MODEL PREDICTIONS OF VIBRIO VULNIFICUS RESPONSE TO CLIMATE VARIABILITY AND CHANGE: A CHESAPEAKE BAY CASE STUDY[11]

ABSTRACT

The effect that climate change and variability will have on waterborne bacteria is a topic of increasing concern for coastal ecosystems, including the Chesapeake Bay. Surface water temperature trends in the Bay indicate a warming pattern of roughly 0.3-0.4^{o}C per decade over the past 30 years. It is unclear what impact future warming will have on pathogens currently found in the Bay, including *Vibrio* spp. Using historical environmental data, combined with three different statistical models of *Vibrio vulnificus* probability, we explore the relationship between environmental change and predicted *Vibrio vulnificus* presence in the upper Chesapeake Bay. We find that the predicted response of *V. vulnificus* probability to high temperatures in the Bay differs systematically between models of differing structure. As existing publicly available datasets are inadequate to determine which model structure is most appropriate, the impact of climatic change on the probability of *V. vulnificus* presence in the Chesapeake Bay remains uncertain. This result points to the challenge of characterizing climate sensitivity of ecological systems in which data are sparse and only statistical models of ecological sensitivity exist.

5.1. Introduction

[11] Urquhart, E.A., Zaitchik B.F., Waugh, D.W., Guikema, S.D., Del Castillo, C.E. Uncertainty in Model Predictions of *Vibrio Vulnificus* Response to Climate Variability and Change: A Chesapeake Bay Case Study. *PLOS ONE*, *(Accepted)*.

Vibrio spp. bacteria are a threat in many coastal aquatic ecosystems around the world (Baker- Austin et al., 2012; Deepanjali et al., 2005; Hendriksen et al., 2011; Cantet et al., 2013; Oberbeckmann et al., 2012). In the Chesapeake Bay, the number of annual human *Vibrio* cases of infection has nearly doubled in the past decade (Maryland Department of Health, 2013; Virginia Department of Health, 2013). Furthermore, *Vibrio spp.* is frequently detected in shellfish harvested for human consumption during the warm summer months (de Magny et al., 2008). In general, this seasonality correlates with peak incidence of *Vibrio* disease caused by *Vibrio spp.* bacteria in many coastal regions (Shapiro et al., 1998; Klontz et al., 1988, Lipp et al., 2002). The probability of finding various *Vibrio* spp. in the Bay varies spatially and seasonally, and researchers have modeled these probability patterns as a statistical function of surface water temperature and salinity (Heidelberg et al., 2002; Jacobs et al., 2010; Wright et al., 1996; Louis et al., 2003; de Magny et al., 2009). These temperature and salinity-based *Vibrio* models have demonstrated skill for available datasets in the Bay and structurally similar statistical models have been applied to predictions of *V. cholerae*, *V. vulnificus*, and *V. parahaemolyticus* in other regions (Oberbeckmann et al., 2012; Eiler et al., 2006; Baker-Austin et al., 2012; Johnson et al., 2012). The environmental range of *V. vulnificus* can vary by region, but in general the bacteria are found in waters with salinity between 5 and 25 and temperature above 15oC (Colwell et al., 1977; Jacobs et al., 2010; Kaper et al., 1981; Lipp et al., 2001).

Recent studies show that surface water temperatures in the Chesapeake Bay have warmed by 0.3-0.4^{o}C per decade over the past 30 years (Austin, 2002; Secor and Wingate, 2008). This trend has resulted in an expansion of the warm season period during which water

temperatures are high enough to support *V. vulnificus* growth: the onset of spring time temperatures (>15^{o}C) has advanced by nearly three weeks (Austin, 2002). Salinity patterns are also sensitive to climate change, as changes in springtime flow of the Susquehanna River - the primary freshwater input to the Bay - can influence salinity throughout the Bay over the *V. vulnificus* growth season. The consensus of climate models is that there will be a rise in winter and spring precipitation in the northern portion of the watershed (Najjar et al., 2009; Hayhoe et al., 2007) implying an increase in January to May Susquehanna River steam flow. A study by Gibson and Najjar (2000) showed that an increase in the January-May Susquehanna stream flow could potentially decrease winter and springtime salinity values by 7% in the upper Chesapeake Bay.

While there is considerable uncertainty in the magnitude of projected warming and freshening of the Chesapeake Bay (Najjar et al., 2010), it would be valuable to understand how a temperature and salinity sensitive pathogen like *V. vulnificus* might respond to observed and projected trends in these environmental parameters. Here we examine three statistical models of *V. vulnificus* probability of presence that demonstrate skill in predicting *V. vulnificus* probability of presence in Chesapeake Bay. All three models use water surface temperature and salinity as the only predictors, but they differ in their structure and in the data used for training and evaluation. Here we evaluate the effect that these differences in structure and training data have on modeled estimates of *V. vulnificus* probability under current climate conditions, which is relevant for pathogen risk assessment and early warning, and consider the implications of these differences for projected *V. vulnificus* risk under climate change.

5.2. Data and Methods

The Chesapeake Bay Estuary, adjacent to the Maryland, Delaware, and Virginia coastline, covers an area of approximately 11,500 km^2 and is characterized by a sharp north-to-south salinity gradient. Salinity ranges from 0-6 in the northern Bay to 18-30 near the mouth of the Bay. Surface water temperatures follow a seasonal cycle, ranging from local wintertime temperatures of -0.5^oC to summertime temperatures of 31^oC (Baird and Ulanowicz, 1989). The Susquehanna River, the largest and northernmost tributary, accounts for roughly 45% of the yearly freshwater inflow into the Bay. This paper focuses on the upper portion of Chesapeake Bay (Fig. 5.1). The upper region of the Bay was selected to avoid model predictions outside of the original training data salinity range (salinity >14).

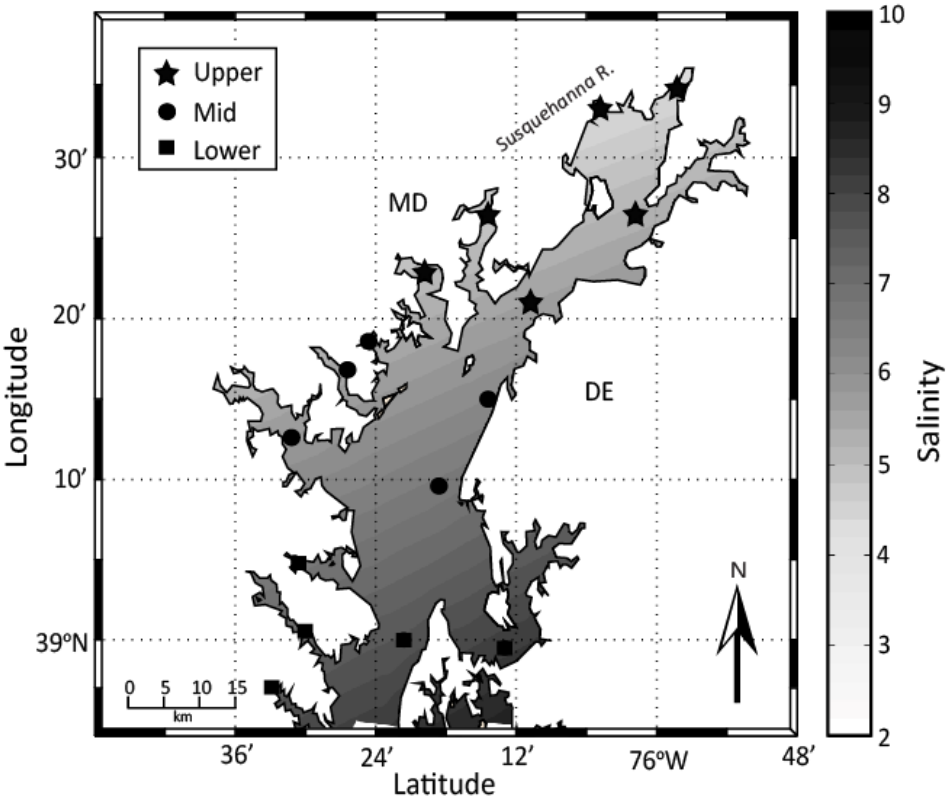

Figure 5.1 Map of the study area, showing contours of average surface water salinity. Dark markers represent in situ monitoring stations used for each of the subregions in this study: upper (star), mid (circle), and lower (square).

The climatological analysis presented here used historical environmental data collected by the Chesapeake Bay Data Program (Chesapeake Bay Program, 2013). Bi-monthly surface water temperature, salinity, and chlorophyll a data were obtained for 16 main stem and tributary monitoring stations (Fig. 5.1) collected from 1985 through 2013. For salinity, the absolute difference between observed salinity and the *V. vulnificus* optimal salinity value of 11.5 (Jacobs et al., 2010) was calculated, and used of deviation from this was used as an explanatory covariate. The 16 monitoring stations were selected based on their geographic location serving as a representation of the upper Chesapeake Bay. In situ data were used to delineate three different salinity zones: upper-upper Bay (hereafter: ”upper region”), middle-upper Bay (hereafter: ”mid region”), and lower-upper Bay (hereafter: ”lower region”). These stations cover the upper main-stem Bay as well as tributary locations, with six stations in the upper region, five stations in the mid region, and five stations in the lower region. Observational data were averaged at monthly intervals for each zone resulting in 337 data records for the upper region and 342 data records for both the mid and lower regions.

These salinity and temperature data were applied to the three statistical *V. vulnificus* probability models available for Chesapeake Bay:

1. NOAA_GLM: The generalized linear model (GLM) of Jacobs et al. (2010): [$z(V.v) =$

$\beta_0 + \beta_1 Temp + \beta_3 |SalnOpt|$, where β_0 is the intercept, β_n is the regression coefficient for the independent covariates, *Temp* is surface temperature, and |*SalnOpt*| is the absolute distance from optimal salinity of 11.5], which was trained using 235 *V. vulnificus* samples collected during the months of July and October of 2007, and April, July, and October of 2008 and were analyzed by the NOAA Chesapeake Bay Office.

2. JHU_GLM: The GLM introduced in Chapter 4 of the same structural form as the NOAA_GLM [$z(V.v) = \beta_0 + \beta_1 Temp + \beta_3|SalnOpt|$] trained using 148 *V. vulnificus*, surface temperature (8-31^{o}C), and surface salinity (0-14) samples collected in the upper Chesapeake Bay during the months of July and September of 2011 and March through June of 2012. Samples were collected by The Johns Hopkins University (JHU) in collaboration with the Maryland Department of the Environment and NASA and were processed at the University of Maryland College Park.

3. JHU_GAM: A generalized additive model (GAM; (Hastie and Tibshirani, 1986)) trained and evaluated using the same data that were used for JHU_GLM: [$z(V.v) = \beta_0 + s_1(Temp) + s_2(Saln)$, where $s_i(x_i)$ is a parameter of the smoothing function, and *Saln* is the salinity value].

We included a GAM in addition to the two structurally identical GLMs because GAM models allow a more flexible regression modeling of the transformed response that combine the predictor variables in a nonparametric manner (Faraway, 2004). All models were implemented in logistic form using a "logit" link function for an optimal prediction point and were trained using observational bacteria data transformed to binary presence/absence. Probability of *V. vulnificus* presence was calculated using $p = e^z /(1 +$

e^z). Diagnostics for each model were performed using Akaike's Information Criterion (AIC) and accuracy (ACC) in an out-of-bag (OOB) cross validation (Breiman L, 1996). ACC is defined as ACC = (TP+TN)/(P+N) where TP is true positive, TN is true negative, P is the number of presence instances, and N is the number of absence instances.

To explore sensitivity of the *V. vulnificus* models to temperature and salinity, we used a range of surface water temperature (0-40^{o}C) and surface salinity (0-13) values as independent model input. Additionally, historical temperature and salinity data were tested as model input, enabling identification of *V. vulnificus* climatology and seasonal trends. To further assess the geographic distribution of the predicted *V. vulnificus* probability for each method, geospatially-interpolated satellite-derived surface temperature and surface salinity (Chapter 2, 3) were used to map spatially complete estimates of probability throughout the upper Bay. Interpolated satellite estimates were developed using monthly, level-2 Moderate Resolution Imaging Spectroradiometer (MODIS) surface water temperature (MOD 28) and ocean color (Rrs 412-678) products.

All statistical computations were carried out in the R Statistical Environment version 2.14, using the 'mgcv' and 'stats' packages, on an Intel Xeon W3580 Processor, 3.33 GHz machine with 12 GB RAM. Computation time for all statistical models was less than one minute.

5.3. Results and Discussion

For model evaluation, goodness of fit and predictive skill for the JHU models were determined using AIC and ACC indices. AIC results indicated that the JHU GAM (145.9) offered better model fit than the JHU GLM (160.4), but performance differences

between models were small relative to measurement uncertainty. NOAA GLM model fit using the NOAA training dataset yielded an AIC of 164.3 (Jacobs et al., 2010). A direct comparison of model fit of could not be calculated due to lack of access to NOAA GLM training data. To predict bacterial presence, selection of an optimal prediction point was required. Rather than setting a prediction point at 0.5 arbitrarily, the prediction point was based on three performance indices: true positive rate, true negative rate, and ACC, yielding an optimal threshold of 0.4 for *V. vulnificus*. To determine the prediction skill of each model, ACC was calculated using the JHU validation dataset (ACC: 0.47, for NOAA GLM, 0.59 for JHU GLM, and 0.60 for JHU GAM). The AIC and ACC values indicated that the JHU models performed significantly better than a null model that only included seasonality as a predictor.

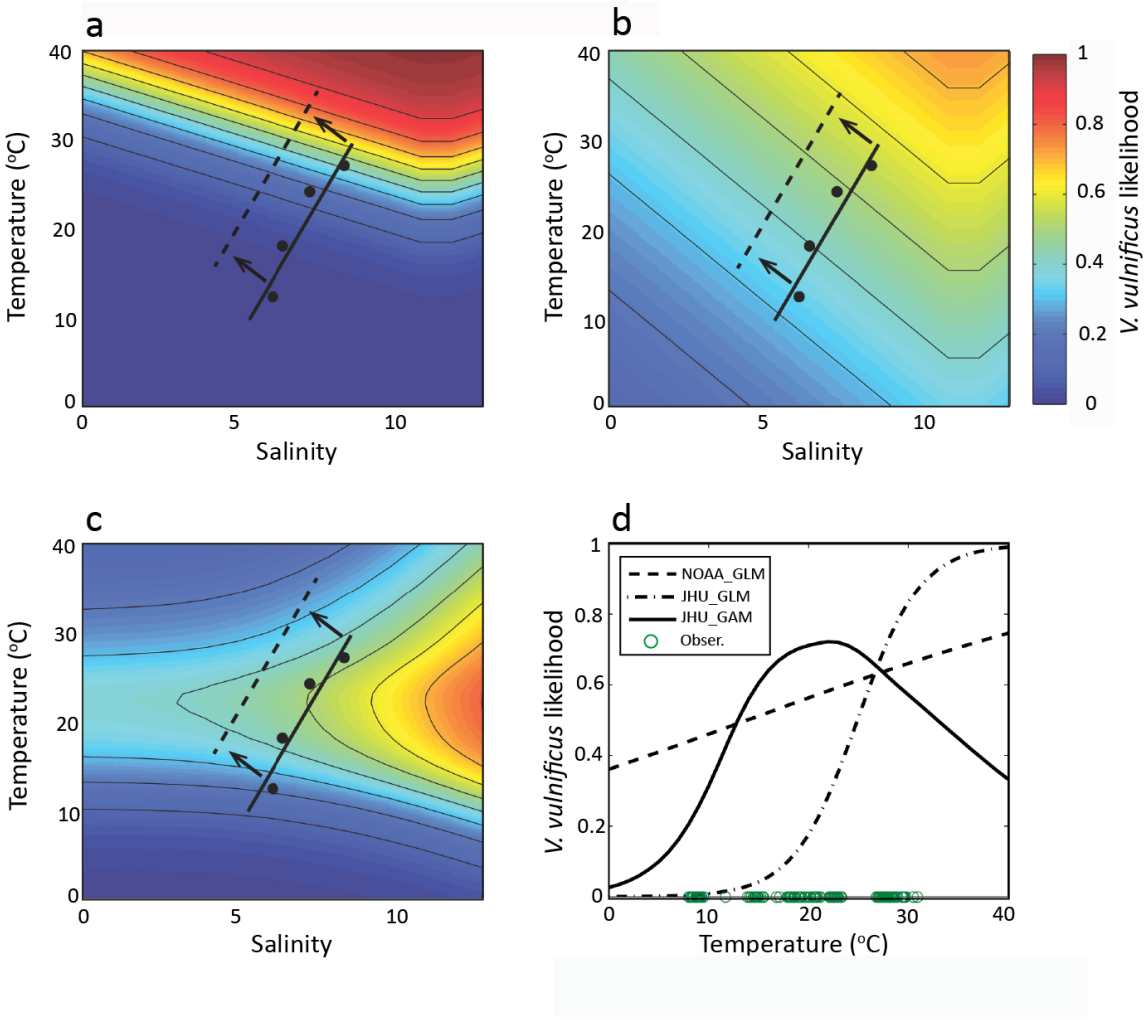

Figure 5.2 Contour plots of *V. vulnificus* probability with temperature and salinity for (a) NOAA GLM, (b) JHU GLM, and (c) JHU GAM. Black dots represent monthly average (April-July) of in situ conditions; black lines represents in situ trend line, and dashed line represents shift in present day temperature and salinity, (d) Plot of temperature regressed against V.vulnificus probability at 11.5 salinity for each empirical method. Green circles represent the range of temperature observations during bacterium sampling.

Figure 5.2 shows the relationship between temperature, salinity, and the mean estimate of predicted *V. vulnificus* probability for each of the tested models, with likelihood levels plotted as contour curves. NOAA GLM (Fig. 2a) exhibits a sharp increase in *V. vulnificus*

probability with increasing temperatures along the axis of optimal salinity (11.5). Similarly, JHU GLM (Fig. 5.2b) exhibits a steady increase in probability with higher temperatures, though the rate of change with temperature is less steep than NOAA GLM. In contrast to the GLMs, JHU GAM (Fig. 5.2c) shows a probability maximum dependent on temperature, indicating a temperature optimum *V. vulnificus* growth above which probability gradually declines. Figure 5.2d offers an alternative view of predicted *V. vulnificus* probability with temperature, at optimal salinity, including temperature observations during in situ bacteria collection. Furthermore, the wide range of observed temperatures confirms that the declining GAM probability above optimal temperature is a valid model response and not an issue of limited observations at high temperature.

These differences in model response also have implications for retrospective or near real-time estimation of risk of *V. vulnificus* presence. Using a 27-year in situ record of temperature and salinity in the upper Chesapeake Bay, we estimated *V. vulnificus* monthly probability of presence according to each statistical model. Fig. 5.3 shows the climatology of surface water temperature and mean estimate model predictions in each region of the upper Bay for March through November. A southward increase in predicted probabilities for all statistical methods during summer months suggests that distance from optimal salinity plays a role in the spatial distribution of *V. vulnificus* presence. Predicted probabilities are likely lower in the upper region due to decreased salinity and larger deviation from optimal salinity. Seasonal patterns in all regions indicate that NOAA_GLM and JHU_GLM predict highest probabilities during the warmest summertime months. JHU_GAM exhibits a bimodal seasonal pattern with peaks in early and late summer across all regions. These JHU_GAM results are consistent with the

temperature dependency shown in Fig. 5.2c.

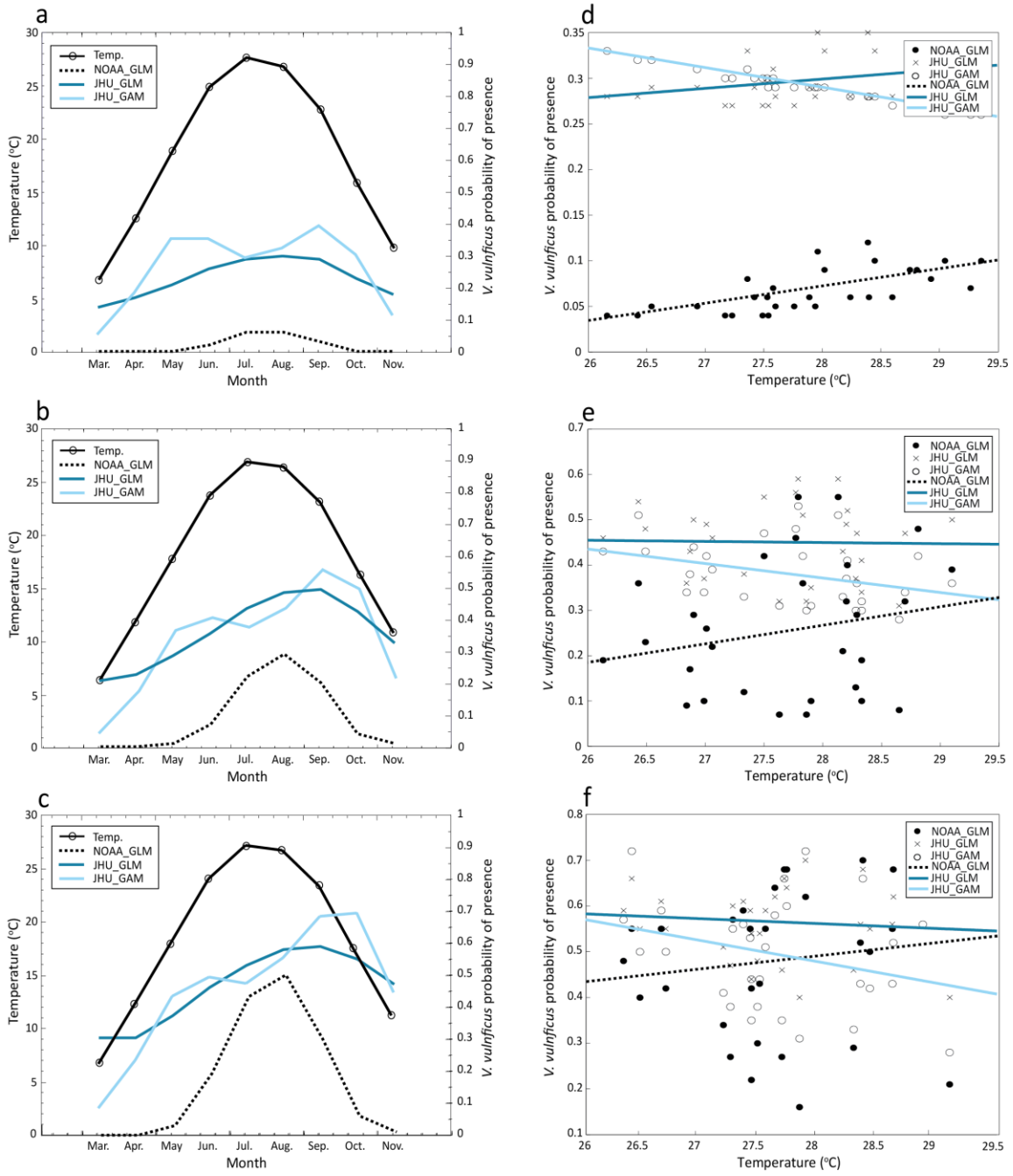

Figure 5.3 Monthly climatology of temperature and *V. vulnificus* probability for each method in the upper (a), mid (b), and lower (c) regions of the Chesapeake Bay. Peak temperature observations by year versus *V. vulnificus* probability for each method in the

upper (d), mid (e), and lower (f) regions of the Chesapeake Bay. Trend lines are included for each method's observations.

The difference in model sensitivity to temperature has implications for characterizing interannual variability in risk. Fig. 5.3d-f show mean predicted *V. vulnificus* probability for the upper, middle, and lower portions of the study area plotted against annual peak monthly SST for the available historical record. In all three subregions, NOAA_GLM predicts that peak probabilities were highest in warmer years, while JHU_GAM predicts the opposite and JHU_GLM falls in between. We emphasize that these are the mean probability estimates for each model, and that there may not be statistically significant differences between model predictions in any given year. Nevertheless, mean estimates are commonly used to communicate risk and to project trends, so the fact that two comparably high performing models – NOAA_GLM and JHU_GAM -- yield opposite mean estimates of the relationship between warm summers and *V. vulnificus* probability is relevant.

The differences in these model response surfaces also have clear implications for projections of *V. vulnificus* probability under climate change. As a simple demonstration, we consider the consensus prediction of warming and freshening of the Bay (dashed line in Figure 5.2 a-c). NOAA_GLM projects steady or increasing probabilities: freshening moves conditions away from the salinity optimum but this effect is offset by increases in predicted probability with rising water temperatures. The JHU_GLM shows a similar pattern but with lower sensitivity to changing environmental conditions. In contrast, warming only increases predicted probability of *V. vulnificus* presence in JHU_GAM for

relatively cool temperatures, representative of spring or fall conditions. Peak summertime temperatures are already above the temperature optimum in this model, so further warming results in a predicted decline in peak summertime *V. vulnificus* probability.

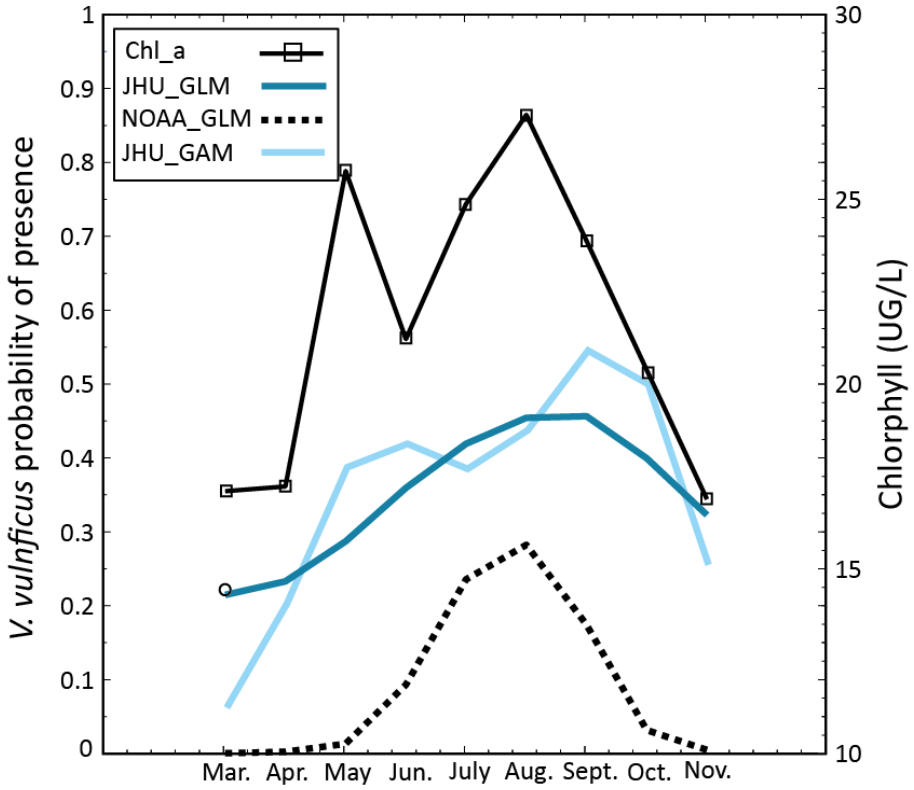

Figure 5.4 Monthly climatology of Chlorophyll a and *V. vulnificus* probability for each method averaged over the entire upper Chesapeake Bay.

While we cannot presently determine which sensitivity pattern is correct—the JHU_GLM and NOAA_GLM increase with higher temperatures or the JHU_GAM decline under warmest conditions—the JHU_GAM behavior might indicate that present-day summertime water temperatures are already above the optimal temperature for *V. vulnificus* growth in Chesapeake Bay. Alternatively, the result might be understood in the

context of previous studies that have shown *Vibrio* dependence on zooplankton due to attachment and/or *Vibrio's* chitinoclastic activity (de Magny et al., 2008; Kaneko and Colwell, 1973). Unfortunately we do not have adequate co-located measurements of zooplankton and *V. vulnificus* to include zooplankton in a predictive model. However, we do find that the climatology of Chesapeake Bay Program in situ chlorophyll a concentrations, which generally correlate with zooplankton presence, exhibits a bimodal seasonal pattern with a slight lead over the JHU GAM predicted *V. vulnificus* peaks (Fig. 5.4).

To examine the geographic extent of each methods' predicted *V. vulnificus* probability, monthly interpolated satellite surface water temperature and surface salinity products were used to create spatially complete probability hind-casts for 2012 in the upper Bay (Fig. 5.5). Consistent with results shown in Fig. 5.3, these maps show highest predicted probability towards the south of the analysis region, where salinity values are closest to optimum. NOAA_GLM and JHU_GLM both show the most widespread zones of high probability in the warmest summer months, while JHU_GAM predicts higher probabilities at the beginning and end of the warm season. Interesting spatial structures are also apparent in these maps. For example, NOAA_GLM predicts slightly elevated *V. vulnificus* probabilities in the eastern waters of the Chesapeake Bay during warmer months, while JHU_GAM predicts high probability zones in the western Bay during months with lower overall probability (Fig. 5.5). These patterns likely reflect the Bay's two-layer physical circulation scheme in which we see fresher surface waters along the western shore and saltier waters along the eastern shore of the Bay. The predictions of statistical *V. vulnificus* probability models compared in this study clearly differ in the

implied relationships between the structure of this circulation and the location of high *V. vulnificus* risk areas.

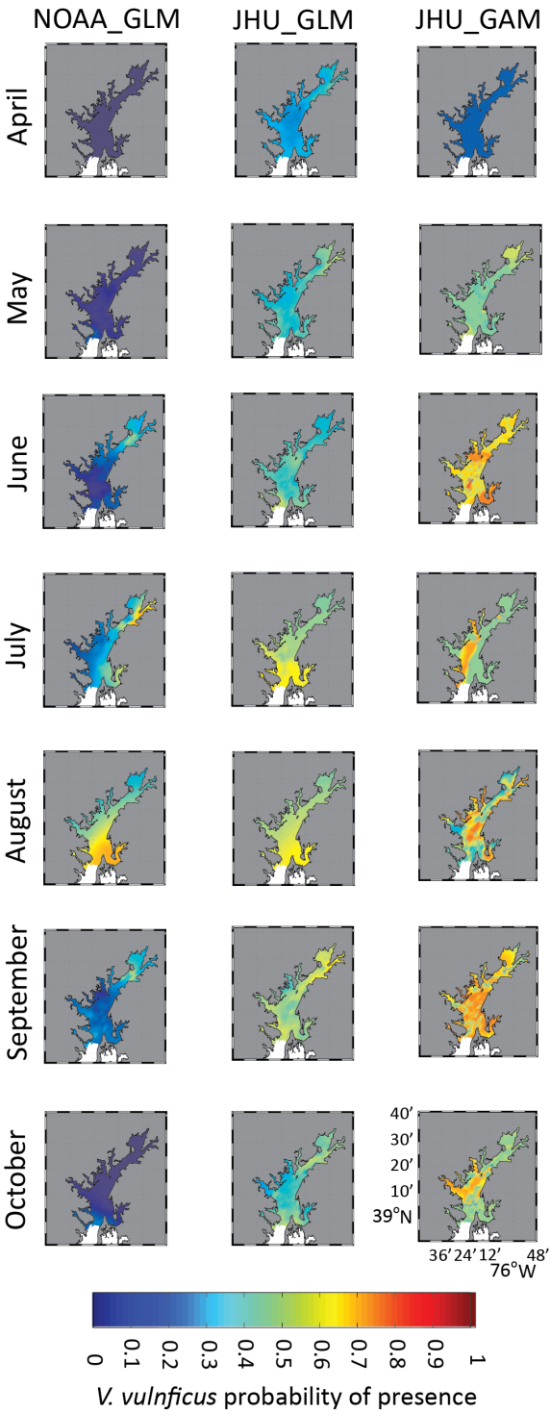

Figure 5.5 Upper Chesapeake Bay monthly *V. vulnificus* probability hind-casts for April through October 2012, for a) NOAA_GLM, b) JHU_GLM, and c) JHU_GAM.

5.4. Conclusions

In summary, this study presents a comparison of three statistical ecological habit models for estimating the probability of *V. vulnificus* presence in the upper Chesapeake Bay. We examined individual model sensitivity to climatic variability and change within the upper Bay by assessing model response to a range of temperature and salinity values. We find that the three models differ systematically in the predicted response of *V. vulnificus* probability to high temperatures in the upper Chesapeake Bay.

These results indicate that more data are required to constrain estimates of climate sensitivity of *V. vulnificus* in Chesapeake Bay: statistical models are limited by the paucity of publicly available data from *V. vulnificus* collections and co-located measurements of ecologically relevant variables, and process-based models would require further research on the *V. vulnificus* life cycle in the Bay. . In addition to different model structure, we acknowledge that the predicted response of *V. vulnificus* likelihood in the Bay may differ between models due to the different spatial and temporal characteristics of the training datasets, as well as different collection and laboratory protocols of each group. Our results also caution against predicting or projecting climate-based changes in *V. vulnificus* exposure risk on the basis of the mean predictions of existing statistical models, as skillful and statistically indistinguishable models differ systematically in predicted *V. vulnificus* sensitivity to rising surface water temperature, even within the range of environmental conditions under which the models were trained.

The challenges facing *V. vulnificus* modeling in Chesapeake Bay are not unique. Indeed, predictive capabilities for climate sensitivity of many pathogens are limited to statistical models based on scarce data. Other recent studies (Ebi, 2008; Hofstra, 2011; Schets et al., 2004) emphasize that the inadequacy of available data hamper climate change projections for a diversity of waterborne pathogen systems in many regions. In the case of *V. vulnificus* in Chesapeake Bay we have a specific example of closely related modeling efforts that suggest systematically different impacts of climate change due to differences in model structure. These kinds of structural comparisons of statistical models, however, are not always performed in studies of climate sensitivity in ecological systems. The results of this study suggest that such model comparisons can be quite important when evaluating uncertainty in climate-based predictions and projections.

6. CHAPTER 6: CONCLUSIONS

It has been suggested that the occurrence of *Vibrio* spp. bacteria is increasing throughout the near shore environments of the Chesapeake Bay. As environmental conditions continue to change in poorly characterized and unpredicted ways, there is a need for more advanced and spatially complete coastal monitoring networks. The thesis was an attempt to model the distribution and occurrence of *Vibrio* spp. bacteria using environmental predictors in the Chesapeake Bay. The intended outcome of this research was to use various forms of environmental data to inform operational and public health risk models for *Vibrio* spp. in shellfish and recreational waters in the Bay. In situ, satellite, interpolated, and modeled estimates of surface water temperature and salinity were used to determine the spatial and temporal distribution and abundance of *Vibrio* spp. in the Chesapeake Bay.

In Chapter 2 I developed and presented the results of multiple statistical models that predict daily, surface salinity across Chesapeake Bay as a function of surface reflectance estimates from NASA MODIS Aqua. Several statistical methods were tested and it was found that surface water salinity could be accurately estimated using remote sensed products with an accuracy that is more than sufficient for many physical and ecological applications in the region. Model evaluation illustrated that a generalized additive model, a generalized linear model, and an artificial neural network performed particularly well in estimating satellite-derived surface salinity in the Chesapeake Bay. Furthermore, Chapter 2 also conducts cross-validation to evaluate the generalizability of the salinity estimates across various temporal, spatial, and fresh water discharge regimes in the Chesapeake

Bay. From the cross-validation results it was concluded that the GAM and GLM outperform the ANN; supporting the original hypothesis that more transparent models can estimate surface water salinity with equal or better accuracy than an artificial neural network.

In the subsequent chapter, I tested several geospatial interpolation techniques as a method for filling spatial data gaps and minimizing errors in the satellite record. Interpolated estimates of satellite-derived surface water temperature and salinity were compared to output of ChesROMS for 2003 and 2007 to assess the relative value of each method for generating inputs into various modeling applications. Results showed that satellite-derived SSS and SST could be geospatially-interpolated with acceptable accuracy in the Bay. In general, the universal kriging method was found to outperform ordinary kriging, and interpolation errors differed systematically from ChesROMS errors both spatially and seasonally in the Chesapeake Bay.

Chapter 4 presents several empirical algorithms for estimating the likelihood of *Vibrio* spp. occurrence and abundance in Chesapeake Bay surface water. To estimate the probability of *Vibrio* spp. being detected in Bay water, I tested several binary classification methods. To model *Vibrio* spp. abundance, several regression methods were applied to samples found positive for *Vibrio* spp. in the Bay. Furthermore, a two-step hybrid approach using a GAM for binary classification and a RF for continuous regression was used to estimate the abundance of *Vibrio* spp. in the absence of previous bacteriological data. Overall, the GAM demonstrated the highest accuracy and improved

positive rate of all the binary models. For *Vibrio* spp. abundance, the RF was found to exhibit lower overall mean absolute error than the other abundance models. Lastly, the hybrid model performed better than the positive only abundance model at sites where *Vibrio* presence had been confirmed by bacteriological methods. The work presented in Chapter 4 offered the ability to predict both presence and abundance of *Vibrio* bacteria in Chesapeake Bay surface water, the later of which is novel to the Chesapeake Bay.

Lastly, Chapter 5 presents a comparison of three statistical ecological habit models introduced in Chapter 4 for estimating the probability of *V. vulnificus* presence in the upper Chesapeake Bay. I examined individual model sensitivity to climatic variability and change within the upper Bay by assessing model response to a range of temperature and salinity values. Model evaluation showed that the three models differed systematically in the predicted response of *V. vulnificus* probability to high temperatures in the upper Chesapeake Bay. Unfortunately, existing publicly available datasets are inadequate to determine which model structure is most appropriate, and thus the impact of climatic change on the probability of *V. vulnificus* presence in the Chesapeake Bay remains uncertain. Ultimately, these results point to the challenge of characterizing climate sensitivity of ecological systems in which data are sparse and only statistical models of ecological sensitivity exist.

6.1 Future Work

The work performed in this thesis has made use of the ever-increasing amount of satellite data, modeled, and in situ environmental parameter data in the Chesapeake Bay.

However, one major challenge of this research has been in the limited data available for the study period. In both the case of in situ modeling of *Vibrio* spp. and model sensitivity to environmental change, it was found that additional bacteria measurements, particularly during warm summer months, are needed to accurately model *Vibrio* spp. and constrain estimates of climate sensitivity in the region. Additional collection and distribution of in situ *Vibrio* measurements is, therefore, critical to future research on this topic. In addition, *Vibrio* modeling could be improved through: hyperspectral remote sensing of environmental data relevant to bacteria occurrence, data merging, sampling and estimation at subsurface depths, and extension of methods to other organisms and coastal regions.

Typically semi-empirical/analytic algorithms developed to estimate concentrations of harmful algal blooms, sea nettles, and other marine organisms in coastal waterbodies have focused on environmental observations from various satellite sensors such as OCM, MODIS, MERIS and Landsat. While the spatial resolutions of some of these satellite sensors are conducive to observations of the larger sections of coastal estuaries like the Chesapeake Bay, the resolutions are too coarse for observation of the smaller upstream rivers and tributaries. Hyperspectral measurements via airborne sensors could provide high-resolution observations ideally suited for detection of plankton and bacteria biomass in small inland systems. Likewise, handheld hyperspectral sensors can provide additional point measurements important for model development and validation studies. Not only does hyperspectral remote sensing enable increased spectral signals, which could be applied to and used to improve the salinity algorithms introduced in Chapter 2, but it also

allows for identification of other optical signatures, such as plankton pigments, that could potentially be useful for *Vibrio* spp. modeling. The ability to monitor harmful bacteria using hyperspectral, high spatial and temporal imagery, could improve the understanding enabling further risk assessment and management strategies in the Chesapeake Bay.

Additionally, as mentioned in Chapter 2 and Chapter 3, to obtain full temporal and spatial coverage of surface water salinity and temperature in the Chesapeake Bay, satellite sensed, interpolated, and in situ parameter observations could be combined with a fluid dynamical model like ChesROMS through data assimilation. Data merging of these environmental observations through the use of numerical modeling could allow for prediction and forecasting of bacteria, and additionally provide full 3 dimensional coverage of the Bay enabling estimates of environmental data into the water column. Previous work (DePaola, et al., 2003; McLaughlin et al., 2005; Thompson et al., 1976) has shown that *Vibrio* spp. bacteria can be isolated from sediment, oysters, and water column samples found in shallow coastal environments. The satellite-derived and interpolated salinity and temperature products discussed in Chapter 2 and 3 are limited to surface waters depths and are thus incapable of estimating and monitoring subsurface *Vibrio* spp. concentration. Data merging though assimilation would enable inference of below surface temperature and salinity conditions that could be used empirically construct estimates of *Vibrio* spp. accumulation in oyster tissues. Not only would this help pinpoint possible high-risk regions for oyster contamination, but could also help to steer the spatial and temporal direction of future field sampling campaigns in the Chesapeake Bay.

Due to a number of factors including but not limited to narrow infrastructure and in-

adequate resources for marine monitoring programs, coastal regions worldwide, including the Chesapeake Bay, currently lack successful harmful marine organism bio-monitoring programs. In light of future projections regarding increasing sea surface temperature and altered salinity conditions in coastal regions, human populations in these regions are increasingly susceptible to risk of pathogenic *Vibrio* disease and harmful algal bloom exposure. However, coastal managers and risk exposure experts in these regions have an advantage in that they have the opportunity to model and improve upon their monitoring programs using examples of successful programs. Looking at the different strengths and weaknesses of existing bio-monitoring programs can be useful in the planning and implementation stages of new and improved marine surveillance programs. One example of a successful marine surveillance system is the California Department of Public Health's (CDPH) Preharvest Shellfish Protection and Marine Biotoxin Monitoring Program employed for harmful algal bloom monitoring off the central coast of California. The CDPH bio-monitoring program is comprised of five basic elements: (1) coastal shellfish and phytoplankton monitoring, (2) monitoring of commercial shellfish product, (3) an annual statewide quarantine on sport-harvested mussels, (4) mandatory reporting of disease cases, and (5) public information and education activities. Extension of these methods to the issue of *Vibrio* monitoring in the Chesapeake Bay could significantly aid in the prediction and, potentially, prevention of human pathogen outbreaks in the region.

REFERENCES

Austin, H. (2002). Decadal oscillations and regime shifts, a characterization of the Chesapeake Bay marine climate. *American Fisheries Society Symposium*, 32, 155-170.

Bahner, L. (2006). User guide for the Chesapeake Bay and tributary interpolator. Annapolis, MD: NOAA, Chesapeake Bay Office.

Baird, D., Ulanowicz, E. (1989). The seasonal dynamics of the Chesapeake Bay Ecosystem. *Ecological Monographs*, 59, 329-364.

Baker-Austin, C., Trinanes, J., Taylor, N., Hartnell, R., Siitonen, A., Martinez-Urtaza, J. (2012). Emerging vibrio risk at high latitudes in response to ocean warming. *Nature Climate Change*, 3, 73-77.

Banakar, V., Constantin de Magny, G., Jacobs, J., Murtugudde, R., Huq, A., Wood, R., Colwell, R. (2011). Temporal and spatial variability in the distribution of Vibrio vulnificus in the Chesapeake Bay: a hindcast study. *Ecohealth*, 8(4), 456-67. doi: 10.1007/s10393-011-0736-4.

Bauer, A., Rorvik, L. (2007). A novel multiplex PCR for the identication of Vibrio parahaemolyticus, Vibrio cholerae and Vibrio vulnificus. *Letters in Applied Microbiology*, 45(4), 371-375. doi: 10.1111/j.1472-765X.2007.02195.x.

Berk, R. (2006). An Introduction to Ensemble Methods for Data Analysis. *Sociological Methods Research*, 34(263).

Blume, H., Fedors, J. (1978). Measurement of ocean temperature and salinity via microwave radiometry. *Bound-Layer Meteorology*, 13, 295-308.

Bowers, D., Brett, H. (2008). The relationship between CDOM and salinity in estuaries:

An analytical and graphical solution. *Journal of Marine Systems*, 73(1-2), 1-7.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. doi: 10.1023/a:1010933404324.

Brown, O. B., Minnett, P. J. (1999). *MODIS Infrared Sea Surface Temperature Algorithm: Algorithm Theoretical Basis Document*. Miami, FL.

Brownlee, K. (1960). Statistical Theory and Methodology in Science and Engineering. John Wiley & Sons, Inc., New York.

Cameron, A. C., Trivedi, P. K. (2013). Regression Analysis of Count Data. Econometric Society Monograph, Cambridge University Press.

Cantet, F., Hervio-Heath, D., Carlo, A., Le Mennec, C., Monteil, C., Quemere, C., Jolivet-Gougeon, A., Colwell, R., Monfort, P. (2013). Quantification of vibrio parahaemolyticus, v. vulnificus, and v. cholera in French Mediterranean coastal lagoons, *Research in Microbiology.*

Centers for Disease Control and Prevention (CDC). Vibrio parahaemolyticus. Retrived January 6, 2013, from http://www.cdc.gov/nczved/divisions/dfbmd/diseases/vibriop/.

Center for Disease Control and Prevention (1998). Outbreak of *Vibrio parahaemolyticus* infections associated with eating raw oysters—Pacific Northwest, 1997. *MMWR. Morbidity and mortality weekly report*, 47, 457-462.

Centers for Disease Control and Prevention. (1999). Outbreak of Vibrio parahaemolyticus infection associated with eating raw oysters and clams harvested from Long Island Sound--Connecticut, New Jersey, and New York, 1998. *MMWR. Morbidity and mortality weekly report*, 48(3), 48.

Chehata, M., Jasinski, D., Monteith, M. C., Samuels, W. B. (2007). Mapping Three-Dimensional Water-Quality Data in the Chesapeake Bay Using Geostatistics1. *JAWRA Journal of the American Water Resources Association*, 43(3), 813-828. doi: 10.1111/j.1752-1688.2007.00065.x.

Chen, R. (1999). In situ fluorescence measurements in coastal waters. *Organic Geochemistry*, 30, 397-409.

Chesapeake Bay Program. (1993). Guide to using Chesapeake Bay Program water quality monitoring data. Annapolis,MD.

Chesapeake Bay Program. (2012). Facts & Figures. Retrieved May 17, 2012, from http://www.chesapeakebay.net/factsandfigures.apsx.

Chesapeake Bay Program. (2013). CBP water quality database (1984-present). Retrieved January 12, 2012, from http://www.chesapeakebay.net/data_waterquality.aspx.

Chipman, H., George, I., McCulloch, R. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4, 266-298.

Colwell, R. R., Kaper, J., Joesph, S. W. (1977). Vibrio cholerae, Vibrio parahaemolyticus, and Other Vibrios: Occurrence and Distribution in Chesapeake Bay. *Science*, 198(4315), 394-396. doi: 10.1126/science.198.4315.394-a.

Committee on Environmental and Natural Resources. (2010). *Scientific Assessment of Hypoxia in U.S. Coastal Waters*. Washington, DC: Interagency Working Group on Harmful Algal Blooms, Hypoxia, and Human Health of the Joint Subcommitee on Ocean Science and Technology.

Constantin de Magny, G., Long, W., Brown, C., Hood, R., Huq, A.. Murtugudde, R., Colwell, R. (2009). Predicting the Distribution of *Vibrio* spp. in the Chesapeake

Bay: A Vibrio cholera case study. *Ecohealth*, 6(3), 378-389. doi: 10.1007/s10393-009-0273-6.

Cressie, N. A. C. (1993). Statistics for Spatial Data. New York: Wiley.

Curriero, F. C. (2006). One the Use of Non-Euclidean Distance Measures in Geostatistics. *Mathematical Geology*, 38(8), 907-926. doi:10.1007/s11004-006-9055-7.

Dayhoff, J., DeLeo, J. (2001). Artificial Neural Networks. *Cancer*, 91, 1615-1635.

Deepanjali, A., Kumar, H., Karunasagar, I. (2005). Seasonal variation in abundance of total and pathogenic vibrio parahaemolyticus bacteria in oysters along the southwest coast of India. *Applied and environmental microbiology*, 71, 3575-3580.

Del Castillo, C., Coble, P. G., Morell, J. M., López, J. M., & Corredor, J. E. et al. (1999). Analysis of the optical properties of the Orinoco River Plume by absorption and fluorescence spectroscopy. *Marine Chemistry*, 66, 35-51.

Del Castillo, C., Coble, P., Morell, J., Lopez, J., Corredor, J. (2001). Multispectral in situ measurements of organic matter and chlorophyll fluorescence in seawater: documenting the intrusion of the Mississippi River plume in the West Florida Shelf. *Limnology and Oceanography*, 46, 1836–1843.

Del Castillo, C. E., Miller, R. L. (2007). On the use of ocean color remote sensing to measure the transport of dissolved organic carbon by the Mississippi River Plume. *Remote Sensing of Environment*, 84(4), 538-549.

Del Vecchio, R., Blough, N. (2004). Spatial and seasonal distribution of chromophoric dissolved organic matter and dissolved organic carbon in the Middle Atlantic Bight.

Marine Chemistry, 89(1-4), 169-187.

DePaola, A., Motes, M. L., Cook, D. W., Veazey, J., Garthright, W. E., Blodgett, R. (1997). Evaluation of an alkaline phosphatase-labeled DNA probe for enumeration of Vibrio vulnificus in Gulf Coast oysters. *Journal of Microbiological Methods*, 29(2), 115-120. doi: 10.1016/S0167-7012(97)00030-4.

DePaola, A., Nordstrom, J. L., Bowers, J. C., Wells, J. G., Cook, D. W. (2003). Seasonal abundance of total and pathogenic Vibrio parahaemolyticus in Alabama oysters. *Applied and environmental microbiology*, 69(3), 1521-1526.

Devore JL. (1995). Probability and Statistics for Engineering and the Sciences. (3rd ed.). Pacific Grove: Brooks/Cole.

Diggle, P. J., Ribeiro, P. J. (2001). geoR: A package for geostatistical analysis. *R-News*, 1(2), ISSN 1609-3631-ISSN 1609-3631.

Diggle, P. J., Ribeiro, P. J. (2007). Model-based Geostatistics (Springer Series in Statistics). New York: Springer.

D'Sa, E., Miller, R. (2003). Bio-optical properties in waters influenced by the Mississippi River during low flood conditions. *Remote Sensing of the Environment*, 84, 538-549.

Ebi, K. (2008). Healthy people 2100: modeling population health impacts of climate change. *Climatic Change*, 88, 5-19.

Eiler, A., Johansson, M., Bertilsson, S. (2006). Environmental influences on vibrio populations in northern temperate and boreal coastal waters (Baltic and Skagerrak seas). *Applied and environmental microbiology*, 72, 6004-6011.

Faraway, J. (2004) Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press.

Fox, J. (2008). Applied Regression Analysis and Generalized Linear Models. SAGE Publications, Inc.

Friedman, J. (1991). Multivariate Adaptive Regression Spline. *The Annals of Statistics*, 19, 1-141.

Geiger, E., Grossi, M., Trembanis, A., Kohut, J., Oliver, M. (2013). Satellite-Derived Coastal Ocean and Estuarine Salinity in the Mid-Atlantic. *Continental Shelf Research*, 63, S235-S242.

Gibson, J., Najjar, R. (2000) The response of Chesapeake Bay salinity to climate induced changes in streamflow. *Limnology and Oceanography*, 45, 1764-1772.

Guikema, S. D. and Quiring, S. M. (2012). Hybrid data mining-regression for infrastructure risk assessment based on zero-inflated data. *Reliability Engineering and System Safety*, 99, 178-182.

Hagy, J. D., Boynton, W. R., Keefe, C. W.,Wood, K. V. (2004). Hypoxia in the Chesapeake Bay, 1950-2001: Long-term change in relation to nutrient loading and river flow. *Estuaries*, 27, 634-658.

Hanley, J. A., McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.

Hastie, T. and Pregibon, D. (1992). Generalized Linear Models in: Statistical Models in S. Chapman and Hall/CRC. London, UK.

Hastie, T., Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, 1, 297-310. doi: 10.1214/ss/1177013604

Hastie, T., Tibshirani, R. (1990). Generalized Additive Models. Chapman and Hall/CRC.

Heidelberg, J., Heidelberg, K., Colwell, R. (2002). Seasonality of Chesapeake Bay

bacteriaoplankton species. *Applied and Environmental Microbiology*, 68, 5488-5497.

Hendriksen, R., Price, L., Schupp, J., Gillece, J., Kaas, R., Engelthaler, D., Bortolaia, V., Pearson, T., Waters, A., Upadhyay, B., et al. (2011). Population genetics of vibrio cholerae from Nepal in 2010: evidence on the origin of the Haitian outbreak. *MBio*, 2.

Hoffman, M. J., Miyoshi, T., Haine, T. W. N., Ide, K., Brown, C. W., Murtugudde, R. (2012). An Advanced Data Assimilation System for the Chesapeake Bay: Performance Evaluation. *Journal of Atmospheric and Oceanic Technology*, 29(10), 1542-1557. doi: 10.1175/jtech-d-11-00126.1.

Hofstra, N. (2011). Quantifying the impact of climate change on enteric waterborne pathogen concentrations in surface water. *Current Opinion in Environmental Sustainability*, 3, 471-479.

Hoge, C. W., Watsky, D., Peeler, R. N., Libonati, J. P., Israel, E., Morris, J. G. (1989). Epidemiology and Spectrum of Vibrio Infections in a Chesapeake Bay Community. *The Journal of Infectious Diseases*, 160(6), 985-993.

Howard, R. J., Bennett, N. T. (1993). Infections caused by halophilic marine Vibrio bacteria. *Annals of Surgery*, 217(5), 525-531.

Hunt, B. R., Kostelich, E. J., Szunyogh, I. (2007). Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter. *Physica D-Nonlinear Phenomena*, 230(1-2), 112-126. doi: 10.1016/j.physd.2006.11.008.

Jacobs, J., Rhodes, M., Brown, C., Hood, R., Leight, A., Long, W., Wood, R. (2010). Predicting the Distribution of Vibrio vulnificus in Chesapeake Bay. NOAA

Technical Memorandum: NOS NCCOS 112(12).

Johnson, C., Bowers, J., Griffitt, K., Molina, V., Clostio, R., Pei, S., Laws, E., Paranjype,. R., Strom, M., Chen, A., et al. (2012) Ecology of vibrio parahaemolyticus and vibrio vulnificus in the coastal and estuarine waters of Louisiana, Maryland, Mississippi, and Washingotn (United States). *Applied and environmental microbiology*, 78, 7249-7257.

Jolliffe, I. T. (2002). Principl Component Analysis (2 ed.). New York: Springer.

Kachan, M., Pimenov, S. (1997). Remote sensing of water salinity at decameter wavelengths. *IEEE Transactions on Geoscience and Remote Sensing*, 35, 302-306.

Kahru, M., Mitchell, B. (2001). Seasonal and nonseasonal variability of satellite-derived chlorophyll and colored dissolved organic matter concentration in the California Current. *Journal of Geophysical Research*, 106, 2517–2529.

Kaneko, T., Colwell, R. (1973). Ecology of Vibrio parahaemolyticus in Chesapeake Bay. *Journal of Bacteriology*, 113(1), 24-32.

Kaneko, T.,Colwell, R. (1974). Incidence of Vibrio parahaemolyticus in Chesapeake Bay. *Applied Microbiology*, 30(2), 251-257.

Kaper, J. B., Remmers, E. F., Lockman, H. and Colwell, R. R. (1981). Distribution of Vibrio parahaemolyticus in Chesapeake Bay during the Summer Season. *Estuaries*, 4(4), 321-327.

Klontz KC, Lieb S, Schreiber M, Janowski HT, Baldy LM, Gunn RA. (1988) Syndromes of *Vibrio vulnificus* infections. Clinical and epidemiologic features in Florida cases, 1981–1987. Ann Intern Med. 109:318-23

Lee, K., Park, J. (1992). Short-term Load Forecasting Using an Artificial Neural Network. *Transactions on Power Systems*, 7(1).

Leifer, I., Lehr, W. J., Simecek-Beatty, D., Bradley, E., Clark, R., Dennison, P., Hu, Y., Matheson, S., Jones, C. E., Holt, B., Reif, M., Roberts, D. A., Svejkovsky, J., Swayze, G., Wozencraft, J. (2012). State of the art satellite and airborne marine oil spill remote sensing: Application to the BP Deepwater Horizon oil spill. *Remote Sensing of Environment*, 124, 185-209.

Lerner, R., Hollinger, J. (1977). Analysis of 1.4 GHz Radiometric measurements from Skylab, *Remote Sensing of the Environment*, 6, 251-269.

Li, R., Kaufman, Y. J., Gao, B., Davis, C. O. (2003). Remote Sensing of Suspended Sediments and Shallow Coastal Waters. *IEEE Transactions on Geoscience and Remote Sensing*, 41(3), 559-566.

Liaw, A., Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2. 18-22

Linz, P., Wang, R. L. (2003). Exploring Numerical Methods: An Introduction to Scientific Computing Using MATLAB. Sudbury, MA: Jones and Bartlett Publishers.

Lipp, E., Rodriguez-Palacios, C., Rose, J. (2001). Occurrence and distribution of the human pathogen Vibrio vulnificus in a subtropical Gulf of Mexico estuary. *Hydrobiologia*, 460(1-3), 165-173. doi: 10.1023/a:1013127517860.

Lipp EK, Huq A, Colwell RR (2002) Effects of global climate on infectious disease: the cholera model. Clinical Microbiology Reviews 15:757–770.

Lorenz, E. N. (1956). Empirical orthogonal functions and statistical weather prediction. Technical report, Statistical Forecast Project Report 1 (49-49). Cambridge, Massachusetts.

Louis, V. R., Russek-Cohen, E., Choopun, N., Rivera, I. N. G., Gangle, B., Jiang, S. C., Rubin, A., Patz, J. A., Huq, A.,Colwell, R. R. (2003). Predictability of Vibrio cholerae in Chesapeake Bay. *Applied and Environmental Microbiology*, 69(5), 2773-2785.

Maisonet, V., Wesson, J., Burrage, D., Howden, S. (2009). Measuring Coastal Sea-Surface Salinity of the Louisiana Shelf from Aerially Observed Ocean Color. Conference proceedings: Oceans 2009 MTS/IEEE. Biloxi, Mississippi

Marlyand Department of the Environment. (2010). Facts about Vibrio Bacteria. Retrieved November 5, 2011, from http://www.mde.maryland.gov/assets/document/vibriofactsheet.pdf.

Maryland Department of Health and Mental Hygiene, Cases of Selcted Notifiable Conditions Reported in Maryland. Retrieved April 6, 2013, from http://phpa.dhmh.maryland.gov/SitePages/disease-conditions-count-rates.aspx/.

Maryland Department of Natural Resources. (2011). Chesapeake Bay Monitoring. Web access: October 4, 2011. http://www.dnr.state.md.us/bay/monitoring/.

McCarthy, S. A., DePaola, A., Cook, D. W., Kaysner, C. A., Hill, W. E. (1999). Evaluation of alkaline phosphatase- and digoxigenin-labelled probes for detection of the thermolabile hemolysin (tlh) gene of Vibrio parahaemolyticus. *Letters in Applied Microbiology*, 28(1), 66-70.

McKeon, J., Rogers, R. (1976). Water quality map of Saginaw Bay from computer processing of Landsat-2 data. Spec. Report to Goddard Space Flight Center, Greenbelt, Maryland.

McLaughlin, J. B., DePaola, A., Bopp, C. A., Martinek, K. A., Napolilli, N. P., Allison, C. G., Middaugh, J. P. (2005). Outbreak of Vibrio parahaemolyticus gastroenteritis associated with Alaskan oysters. *New England Journal of Medicine*, *353*(14), 1463-1470.

Morel, A., Gentili, B. (2009). A simple band ratio technique to quantify the colored dissolved and detrital organic material ocean color remotely sensed data. *Remote Sensing of the Environment*, 113, 998-1011.

Morris, J., Black, R. (1985). Cholera and other vibrioses in the United States. *New England Journal of Medicine*, 312, 343-350.

Motes, M. L., DePaola, A., Cook, D. W., Veazey, J. E., Hunscuker, J. C., Garthright, W. E., Blodgett, R. J., Chirtel, S. J. (1998). Influence of Water Temperature and Salinity on Vibrio vulnificus in Northern Gulf and Atlantic Coast Oysters (Crassostrea virginica). *Applied and Environmental Microbiology*, 64(4), 1459-1465.

Murphy, R. R., Curriero, F. C., Ball, W. P. (2010). Comparison of Spatial Interpolation Methods for Water Quality Evaluation in the Chesapeake Bay. *Journal of Environmental Engineering*, 136(2), 160-171. doi: 10.1061/(ASCE)EE.1943-7870.0000121.

Murphy, R. R., Perlman, E., Ball, W. P., Curriero, F. C. (2012). Kriging with Water Distance in Chesapeake Bay. *(In Prep.)*

Najjar, R., Patterson, L, Graham, S. (2009). Climate simulations of major estuarine watersheds in the mid-Atlantic region of the US. *Climatic Change*, 95, 139-168.

Najjar, R., Pyke, C., Adams, M., Breitburg, D., Hershner, C., Kemp, M., Howarth, R., Mulholland, M., Paolisso, M., Secor, D., et al. (2010) Potential climate-change impacts on the Chesapeake Bay. *Estuarine, Coastal and Shelf Science*, 86, 1-20.

Nateghi, R., Guikema, S. D. (2013). Estimating Power Distribution System Outages during Tropical Cyclones in the Gulf Region of the U.S. with Reduced Complexity Models. *Risk Analysis*, *(under review).*

National Oceanic and Atmospheric Administration (NOAA). (2010). Remote Sensing for Coastal Management. Sea Nettle Forecast. CoastWatch Program, Chesapeake Bay office. Retrieved on October 4, 2011, from http://chesapeakebay.noaa.gov/forecasting-sea-nettles.

National Aeronautical Space Administration (NASA). (2011). MODIS Information Page. Retrieved on May 1, 2011, from http://modis.gsfc.nasa.gov/.

Nelder, J. A., Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*, *Series A* (General) 135(3), 370-384.

Nelson, E., Harris, J., Glenn Morris, J., Calderwood, S., Camilli, A. (2009). Cholera transmission: the host, pathogen and bacteriophage dynamic. *Nature Reviews Microbiology*, 7, 693-702

Oberbeckmann, S., Fuchs, B., Meiners, M., Wichels, A., Wiltshire, K., Gerdts, G. (2012). Seasonal dynamics and modeling of a vibrio community in coastal waters of the North Sea. *Microbial ecology*, 63, 543-551.

Ondrusek, M., Stengel, E., Kinkade, C. S., Vogel, R. L., Keegstra, P., Hunter, C., Kim, C. (2012). The development of a new optical total suspended matter algorithm for the Chesapeake Bay. *Remote Sensing of Environment*, 119, 243-254.

Ortiz, D. M., Tissot, B. N. (2008). Ontogenetic patterns of habitat use by reef-fish in a Marine Protected Area network: a multi-scaled remote sensing and in situ approach. *Marine Ecology Press Series*, 365, 217-232.

Parveen, S., DaSilva, L., DePaola, A., Bowers, J., White, C., Munasinghe, K. A., Brohawn, K., Mudoh, M., Tamplin, M. (2013). Development and validation of a predictive model for the growth of Vibrio parahaemolyticus in post-harvest shellstock oysters. *International Journal of Food Microbiology*, 161(1), 1-6. doi: 10.1016/j.ijfoodmicro.2012.11.010.

Parveen, S., Hettiarachchi, K. A., Bowers, J. C., Jones, J. L., Tamplin, M. L., McKay, R., Beatty, W., Brohawn, K., DaSilva, L. V., DePaola, A. (2008). Seasonal distribution of total and pathogenic Vibrio parahaemolyticus in Chesapeake Bay oysters and waters. *International Journal of Food Microbiology*, 128(2), 354-361. doi: 10.1016/j.ijfoodmicro.2008.09.019.

Prasad, M., Long, W., Zhang, X., Wood, R. J., Murtugudde, R. (2011). Predicting dissolved oxygen in the Chesapeake Bay: applications and implications. *Aquatic Sciences-Research Across Boundaries*, *73*(3), 437-451.

Preisendorfer, R. W. (1988). Principal Component Analysis in Meterology and Oceanography. New York: Elsevier.

Pritchard, D. (1952). Salinity Distribution and circulation in the Chesapeake estuarine system. *Journal of Marine Research*, 11, 106-123.

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. http://www.R-project.org.

Schabenberger, O., Gotway, C. (2004). Statistical methods for spatial data analysis. Boca Raton, FL.: CRC.

Schets, F., Engels, G., Evers, E. (2004). Cryptosporidium and giardia in swimming pools in the Netherlands. *J Water Heatlh*, 2, 191-200.

Sheffield, J., Wood, E., Anderson, B., Bradbury, J., DeGaetano, A., et al. (2007). Past and future changes in climate and hydrological indicators in the US northeast. *Climate Dynamics*, 28, 381-407.

Secor, D., Wingate, R. A 69-year record of warming in the Chesapeake Bay. *Fisheries*, (*in review).*

Shapiro RL, Altekruse S, Hutwagner L, Bishop R, Hammond R, Wilson S, et al. (1998) The role of Gulf Coast oysters harvested in warmer months in *Vibrio vulnificus* infections in the United States, 1988–1996. *Vibrio* Working Group. The Journal of Infectious Diseases 178:752–759.

Stergiou, C., Siganos, D. (1996). Neural Networks. Web Training Doc. Retrieved on May 12, 2011, from http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html.

Strom, M., Paranjpye, R. (2000). Epidemiology and pathogenesis of Vibrio vulnificus. *Microbes Infection*, 2(2), 177-188. doi: 10.1016/S1286-4579(00)00270-7.

Sutton, C. (2005). Classification and Regression Trees, Bagging, and Boosting. Handbook of Statistics: Data Mining and Data Visualization. *(*Volume 24). Penn State University, PA.

The MathWorks Inc. (2010). MATLAB (Version R2010a). Natick, Massachusetts.

Thompson, C. A., Vanderzant, C. (1976). Relationship of Vibrio parahaemolyticus in oysters, water and sediment, and bacteriological and environmental indices. *Journal of Food Science*, 41(1), 117-122.

United States Geological Survey. (2012). USGS Real-Time Water Data for the Nation. USGS 01578310 Susquehanna River at Conowingo, MD. Retrieved on February 10, 2012, from http://waterdata.usgs.gov/usa/nwis/uv?01578310.

Urquhart E., Hoffman M., Zaitchik B, Guikema S., Geiger E. (2012) Remotely Sensed Estimates of Surface Salinity in the Chesapeake Bay: A Statistical Approach. *Remote Sensing of Environment*, 123: 522-531.

Urquhart E., Hoffman M., Murphy R., Zaitchik B. (2013) Geospatial Interpolation of MODIS-Derived Salinity and Temperature in the Chesapeake Bay. *Remote Sensing of Environment.* 135: 167-177.

Urquhart E.A., Guikema S.D., Zaitchik B.F., Haley B.J., Taviani, E., Chen, A., Brown, M.E., Huq, A., Colwell, R.R. Use of Environmental Parameters to Model Pathogenic *Vibrios* in Chesapeake Bay. *Journal of Environmental Informatics*. (*Accepted).*

Urquhart, E.A., Zaitchik B.F., Waugh, D.W., Guikema, S.D., Del Castillo, C.E. (2014) Uncertainty in Model Predictions of *Vibrio Vulnificus* Response to Climate Variability and Change: A Chesapeake Bay Case Study. *PLOS ONE*, *(Accepted).*

Virginia Departmental Health, Virginia Reportable Disease Surveillance Data. Retreived on May 5, 2013, from http://www.vdh.virginia.gov/Epidemiology/Surveillance/SurveillanceData/.

Water Encyclopedia. (2009). Water: Chesapeake Bay. Retrieved on October 12, 2011 from http://www.waterencyclopedia.com/Ce-Cr/Chesapeake-Bay.html.

Wilks, D. (2006). Principal Component (EOF) Analysis (pp. 463-507): Elsevier Inc.

Wood, S. N. (2006). Generalized Additive Models: an introduction with R. Taylor & Francis Group; Chapman & Hall/CRC.

World Health Organization (WHO). (2005). Risk assessment of Vibrio vunlificus in raw oysters. MRA. Microbiological Risk Asssessment Series. 8.

World Health Organization (WHO). (2013). Weekly epidemiolgical record (WER). 88(31), 321-336.

Wright, A. C., Hill, R. T., Johnson, J. A., Roghman, M. C., Colwell, R. R., Morris, J. G. (1996). Distribution of Vibrio vulnificus in the Chesapeake Bay. *Applied and Environmental Microbiology*, 62(2), 717-724.

Xu, J., Chao, S.-Y., Hood, R. R., Wang, H. V., Boicourt, W. C. (2002). Assimilating high-resolution salinity data into a model of a partially mixed estuary. *Journal of Geophysical Research*, 107(3074), 14.

Xu, J., Long, W., Wiggert, J., Lanerolle, L. J., Brown, C., Murtugudde, R., Hood, R. (2012). Climate Forcing and Salinity Variability in Chesapeake Bay, USA. *Estuaries and Coasts*, 35(1), 237-261. doi: 10.1007/s12237-011-9423-5.

Yamazaki, K. and Nwadiuto, E. (2012). Environmental Predictors of Pathogenic Vibrios in South Florida Coastal Waters. *The Open Epidemiology*, 5, 1-4.

CURRICULUM VITAE

Erin A. Urquhart
Born May 26, 1985 in Canoga Park, CA

EDUCATION

2009-4/2014	Ph.D. Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD. Dissertation: *Remote Sensing and Modeling of Vibrio Bacteria in the Chesapeake Bay.* Thesis Advisors: Dr. Benjamin Zaitchik
2011	M.A. Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD. Thesis Advisors: Dr. Benjamin Zaitchik and Dr. Darryn Waugh
2009	M.H.S. Environmental Health, Johns Hopkins School of Public Health, Baltimore, MD. Thesis: *Analysis of Marine Bio-toxins, Climate Change, and Coastal Monitoring Program*. Thesis Advisors: Dr. Thaddeus Graczyk and Dr. Jacqueline Agnew
2007	B.A. Global Environmental Studies, Sonoma State University, Rohnert Park, CA. Minor: Geography

PROFESSIONAL EXPERIENCE

2009-Present	Graduate Research Assistant/Teaching Assistant, Johns Hopkins University**,** Baltimore, MD.
2007	Research Assistant, Marine Bio-toxin Monitoring Program. California Department of Public Health/University of California Santa Cruz, San Luis Obispo, CA.

FIELD EXPERIENCE

2011-2012	R/V Kerhin. Water and bacterial collection. Chesapeake Bay**,** MD.
2009	Insect collection for Lyme disease survey. Baltimore County, MD.

FELLOWSHIPS AND HONORS

2013	Student Oral Presentation Competition, MEDGEO/ISPRS 2013 Annual Meeting, Arlington, VA
2012	Outstanding Student Presentation Award, AGU Ocean Sciences Meeting Salt Lake City, UT.
2011-2012	Graduate Student Summer Field Work Grant. Earth and Planetary Sciences, Johns Hopkins University
2010	Graduate Student Fellowship/Seed Grant, Global Water Program, Johns Hopkins University

RESEARCH INTERESTS

Coastal and aquatic environments - Climate impacts on coastal systems - Environmental health - Ecological modeling - Geospatial statistics - Operational applications of coastal remote sensing - Marine pathogens, bacteria, and harmful algal blooms - Risk assessment - Water quality monitoring, mapping and modeling - Coastal hazards - Empirical model development

REFEREED PUBLICATIONS

Urquhart, E.A., Zaitchik B.F., Waugh, D.W., Guikema, S.D., Del Castillo, C.E. (2014)Uncertainty in Model Predictions of *Vibrio Vulnificus* Response to Climate Variability and Change: A Chesapeake Bay Case Study. *PLOS ONE*. *(Accepted).*

Urquhart E.A., Guikema S.D., Zaitchik B.F., Haley B.J., Taviani, E., Chen, A., Brown, M.E., Huq, A., Colwell, R.R. (2014) Use of Environmental Parameters to Model Pathogenic *Vibrios* in Chesapeake Bay. *Journal of Environmental Informatics. (Accepted).*

Urquhart E.A., Hoffman M.J., Murphy R.R., Zaitchik B.F. (2013) Geospatial Interpolation of MODIS-Derived Salinity and Temperature in the Chesapeake Bay. *Remote Sensing of Environment.* 135: 167-177.

Urquhart E.A., Hoffman M.J, Zaitchik B.F, Guikema S.D., Geiger E.F. (2012) Remotely Sensed Estimates of Surface Salinity in the Chesapeake Bay: A Statistical Approach. *Remote Sensing of Environment.* 123: 522-531.

CONFERENCE & INVITED PRESENTATIONS

Urquhart, E.A. (2014) Oral presentation: Use of Environmental Predictors to Model Pathogenic *Vibrios* in Chesapeake Bay. 2014 American Meteorological Society. February 2-6. Atlanta, GA.

Urquhart, E.A. (2013) Invited talk: Use of Remotely Sensed Predictors to Model *Vibrios* in Chesapeake Bay, MD. University of New Hampshire. November 15.

Urquhart, E.A. et al., (2013) Oral presentation: Remote Sensing of *Vibrio* spp. Bacteria in Chesapeake Bay Estuary, MD. 2013 Conference of the International Medical Geology Association, ISPRS: Advances in Geospatial Technologies for Health II. August 25-29. Arlington, VA.

Urquhart, E.A. et al., (2013) Oral presentation: Spatial Interpolation of Satellite-derived Temperature and Salinity in the Chesapeake Bay: an Ecological Forecasting Application. ASPRS 2013 Annual Conference. March 24-28. Baltimore, MD.

Urquhart, E.A. et al., (2012) Oral presentation: Geospatial Interpolation of Remotely Sensed Observations in the Chesapeake Bay: an Ecological Forecasting Application. AGU Fall Meeting. December 3-7. San Francisco, CA.

Urquhart, E.A. et al., (2012) Poster presentation: Remotely Sensed Estimates of Surface Salinity and Environmental Vibrio in the Chesapeake Bay. Ocean Optics 2012, October 8-11. Glasgow, Scotland.

Urquhart, E.A. et al., (2012) Oral presentation: Remotely Sensed Estimates of Surface Salinity and Environmental Vibrio in the Chesapeake Bay. Chesapeake Bay Modeling Symposium. May 21-22. Annapolis, MD.

Urquhart, E.A. et al., (2012) Oral presentation: Remotely Sensed Estimates of Surface Salinity and Environmental Vibrio in the Chesapeake Bay. AGU Ocean Sciences Meeting. February 19-24. Salt Lake City, UT.

Urquhart, E.A. et al., (2011) Oral presentation: Satellite Remote Sensing of Environmental Vibrio in the Chesapeake Bay. ISPRS: Advances in Geospatial Technologies for Health. September 11-14. Santa Fe, NM.

WORKSHOPS

OSU/NASA Dream Ocean Satellite Image Workshop; Fisheries & Aquaculture (2013) June 4 -5. Newport, OR.

PROFESSIONAL ACTIVITIES

Journal Reviewer: Remote Sensing of Environment, IEEE Geoscience and Remote Sensing Letters

PROFESSIONAL MEMBERSHIP

2010-Present Member, American Geophysical Union
2012-Present Member, Association for Limnology and Oceanography